FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Using Data Mining to Predict Students' Academic Success

André Filipe Roque Silva

DISSERTATION

Mestrado Integrado em Engenharia Informática e Computação

Supervisor: Vera Lucia Miguéis Oliveira e Silva

February 28, 2016

Using Data Mining to Predict Students' Academic Success

André Filipe Roque Silva

Mestrado Integrado em Engenharia Informática e Computação

Approved in oral examination by the committee:

Chair: Doctor João Pedro Mendes Moreira
External Examiner: Doctor Paulo Alexandre Ribeiro Cortez
Supervisor: Doctor Vera Lucia Miguéis Oliveira e Silva

February 28, 2016

Abstract

Currently, universities record large amounts of data about students. Despite its potential to inform decisions regarding the allocation of resources and efforts, this information tends to be overlooked. Educational data mining is a recent research field that focuses on the use of data mining techniques to transform large volumes of educational data into useful and relevant knowledge that can improve the educational processes and decisions.

This work intends to propose a set of three models. The first two will use the information available at start of the first semester and second semester, respectively, of the first year of the student's academic path to predict the academic success of the students enrolled at FEUP at the end of that semester. The third model will use the information available at the end of the first year to predict the academic performance of the students enrolled at FEUP at the end of their degree. At the same time, this work also intends to identify the factors that are the most critical to these models.

The results of this project could allow college principals to identify students in need of more pedagogical support, as well as students with a high probability of excelling in their studies. It could also allow them to focus their attention on the critical aspects, by implementing mechanisms that tackle students' difficulties.

The first step of the developed work consists of data cleaning and preparation processes that normalize the data retrieved from the university's information system and the definition of the derived variables. The second step is an empirical analysis of different algorithms with interpretable models in order to identify the models best suited for the tasks of predicting academic success regarding their performance with the available entry data. Thirdly, an analysis of the generated models will be presented, along with the identification of their main predictive attributes.

Resumo

Actualmente, as universidades acumulam grandes quantidades de dados sobre os seus estudantes. Apesar do seu potencial para informar decisões sobre a alocação de recursos e esforços, esta informação tende a ser ignorada. Data Mining Educacional é uma área de investigação recente que se foca no uso de técnicas de data mining para transformar grandes volumes de dados educacionais em conhecimento útil e relevante que possa melhorar as decisões e os processos educativos.

Este projecto pretende propôr um conjunto de três modelos. Os primeiros dois modelos irão utilizar informação disponível no início do primeiro e segundo semestres, respectivamente, do primeiro ano do percurso académico do estudante para prever o sucesso académico de alunos da FEUP no final desse semestre. O terceiro modelo irá utilizar os dados disponíveis ao final do primeiro ano para prever o desempenho académico dos estudantes da FEUP aquando do final do seu curso. Ao mesmo tempo, este projecto também pretende identificar os factores com maior poder discriminativo nestes modelos.

Os resultados deste projecto poderão permitir aos directores das faculdades identificar estudantes com necessidade de apoio pedagógico, assim como aqueles com uma alta probabilidade de obter resultados excelentes. Também poderão permitir que foquem a sua atenção nos aspectos mais críticos, através da implementação de mecanismos que combatam as dificuldades dos estudantes.

O primeiro passo do projecto desenvolvido consiste em processos de limpeza e preparação de dados que normalizam os dados recolhidos do sistema de informação da universidade e na definição das variáveis derivadas. O segundo passo é uma análise empírica dos diferentes algoritmos com modelos interpretáveis de forma a identificar os modelos mais apropriados às tarefas de prever sucesso e desempenho académico com os dados de entrada disponíveis. Por último, serão apresentadas análises dos modelos gerados, assim como a identificação dos seus principais atributos predictivos.

Acknowledgements

First, I would like to thank my supervisor, Vera Lucia Miguéis Oliveira e Silva, PhD., for always being available and for all the support she gave me during the development of this dissertation.

Secondly, I would also like to thank LEA and FEUP for the opportunity to engage in this project, for making the administrative data I required for the development of the prediction models available to me, and for generally providing me with all the conditions I required to successfully complete this project.

Lastly, I want to thank Cristiano Seabra for all the support and companionship during the development of this project, and for keeping me motivated to always demand more of myself.

André Silva

"We are drowning in information but starved for knowledge."

John Naisbitt

Contents

List of Figures

List of Tables

Abbreviations

A3ES	Agency for Assessment and Accreditation of Higher Education
AUC	Area Under Curve
ECTS	European Credit Transfer System
EDM	Educational Data Mining
FEUP	Faculdade de Engenharia da Universidade do Porto
GPA	Grade Point Average
KKT	Karush–Kuhn–Tucker
LEA	Laboratório de Ensino e Aprendizagem
LEC	Licenciatura em Engenharia Civil
LEEC	Licenciatura em Engenharia Electrotécnica e de Computadores
LEGA	Licenciatura em Engenharia e Gestão do Ambiente
LEIC	Licenciatura em Engenharia Informática e Computação
LEM	Licenciatura em Engenharia Mecânica
LEMG	Licenciatura em Engenharia de Minas e Geoambiente
LEMM	Licenciatura em Engenharia Metalúrgica e de Materiais
LEQ	Licenciatura em Engenharia Química
LGEI	Licenciatura em Gestão e Engenharia Industrial
MIB	Mestrado Integrado em Bioengenharia
MIEA	Mestrado Integrado em Engenharia do Ambiente
MIEC	Mestrado Integrado em Engenharia Civil
MIEEC	Mestrado Integrado em Engenharia Electrotécnica e de Computadores
MIEIC	Mestrado Integrado em Engenharia Informática e Computação
MIEIG	Mestrado Integrado em Engenharia Industrial e Gestão
MIEM	Mestrado Integrado em Engenharia Mecânica
MIEMM	Mestrado Integrado em Engenharia Metalúrgica e de Materiais
MIEQ	Mestrado Integrado em Engenharia Química
ROC	Receiver Operating Characteristic
SMO	Sequential Minimal Optimization
SVM	Support Vector Machine

Chapter 1

Introduction

1.1 Context

Education plays a crucial role in a country's development, both in terms of its individual citizens and of its society as a whole. One of the goals established by the European Union for 2020 stipulates the attainment of tertiary education by at least 40% of its population aged between 30 and 34 years.

More specifically in Portugal, this lead to the creation of the Agency for Assessment and Accreditation of Higher Education (A3ES). This Agency promotes the creation of tools that contribute to the increase of the graduation rate in Higher Learning Institutions. At Faculdade de Engenharia da Universidade do Porto, LEA was created in order to internally promote initiatives and the creation of tools that improve academic success and the quality of learning.

The most recent tools adopted by academic institutions make use of Information Systems, as these have become prevalent as key resources for their administrative applications.

The amount of data in these systems marks them as a valuable source of knowledge with regards to understanding students and their environment, which could lead to improvements in the quality of learning.

1.2 Motivation and Goals

The resources available to higher learning institutions are limited, therefore it is important that they are not wasted. Identifying students who are expected to not achieve academic success very early, especially within the first year of their curricula, would allow the managers of these institutions to direct support to those students.

Therefore, this dissertation will focus on producing models that accurately identify students who are expected to not achieve academic success on both semesters of the first year, as well as one that predicts long term academic performance after the first year. At the same time, these models will be analyzed so as to identify the factors that most affect their predictions.

1.3 Structure of the Dissertation

Besides this introductory chapter, this dissertation contains 6 more chapters. In chapter 2 we review the existing literature on the topic and present some related works. In chapter 3, the methodology utilized in this project is presented. After that, in chapter 4 we present an exploratory analysis of the data and discuss the possible implications. We also give an overview of the data preparation tasks performed. This is followed, in chapter 5, by the results of the evaluation of the implemented models. Lastly, in chapter 6 we discuss the results presented in the previous chapter, as well as further avenues of analysis.

Chapter 2

Literature Review

In this chapter we review the existing literature on Educational Data Mining and Academic Success. We begin by giving an overview of these subjects and then focusing on challenges identified by other authors that might affect this project, as well as on the solutions those authors present for overcoming those issues.

2.1 Educational Data Mining

Educational Data Mining (EDM) concerns itself with the use of Data Mining techniques in the aforementioned Information Systems, tapping into the data contained therein in order to extract meaningful knowledge that can support the decision-making process by enabling a better understanding of students and their learning environment (R. S. Baker & Yacef, 2009).

Romero and Ventura (2007) and Baker and Yacef (2009) review the research in EDM on the periods of 1995 to 2005 and 2005 to 2009, respectively. From these reviews we are able to identify the main areas of EDM as being:

- **Student attrition**: this revolves around identifying the main factors that lead to students attaining graduation or not. EDM has been successful in this area, as can be seen in the work of Pedro Strecht, J. M. Moreira and C. Soares (2014), however the results are usually specific to the data set used.

- **Improvement of student models**: Baker and Yacef (2009) define student models as a representation of a students' characteristics or his state, as well as his knowledge level, motivation, meta-cognition and behaviors. These models allow educational software systems to adapt their responses to the student. This area focuses on improving the models through an increase in the student attributes that are analyzed or through the integration of high-level constructs such as the experience of poor self-efficacy on the part of the student. Research in this area by Baker et al. (2006) focused on the identification of students with feelings of frustration.

- **Personalized learning environments and recommendation systems**: personalized learning environments are learning systems which adapt themselves to a students' characteristics. These are closely related to recommendation systems, which can recommend lines of inquiry to a student to allow him to further his educational goals. As referred by Huebner (2013), however, the need for these recommendations to be in line with the overall educational goals is what differentiates it from the broadly used commercial recommendation systems.

- **Improvement of resource management systems**: this concerns the integration of Data Mining tools and the work developed in the other areas of EDM into the existing management systems, with a focus on maintaining high levels of usability so as to allow average users to take advantage of such tools. It does not cover the development of the tools themselves or their use in further understanding the learning environment, focusing instead on the usability and the ease of use of the integrated interfaces with the existing systems. An example of such an implementation can be found in the work of Garcia et al. (2011).

- **Study of pedagogical support**: this revolves around identifying the most effective type of pedagogical support for a given situation and group of students. Beck and Mostow (2008) associated a student's performance with the amount of each type of pedagogical support that he received.

- **Study of educational theories**: this area concerns itself with the empirical analysis of educational theories and phenomenons, enabling a deeper comprehensions of the key aspects of these theories (R. S. Baker & Yacef, 2009). With regards to this, Gong et al. (2009) analyzed the impact of self-discipline in learning, showing that despite there being a correlation between self-discipline and the number of mistakes made, its real impact in learning was minimum.

Furthermore, work in these areas can also be classified with regards to the task performed. Both Romero and Ventura (2007) and Baker and Yacef (2009) present their own taxonomy of these tasks, with the one presented by Romero and Ventura being more directed toward EDM in web data and the one by Baker and Yacef being more general. From the analysis of both taxonomies, we can list the tasks as considered by both authors as being:

- **Prediction**: the determination of the value of a variable through other variables whose values are known, called predictors. Prediction problems may be divided into classification problems, where the objective is to classify the unknown variable as belonging to one of several pre-established classes, and regression problems, where the objective is to predict the value of a continuous numerical variable (Kotsiantis, Zaharakis, & Pintelas, 2007). Nghe, Janecek and Haddawy (2007) and Kabakchieva (2013) use several different prediction algorithms to predict student performance in their work, such as decision trees and Bayesian classifiers.

- **Outlier Detection**: according to Hodge and Austin (2004), an outlier is an instance which appears to be inconsistent with the remainder of the data set it belongs to. In the context of EDM, detecting such outliers allows for the identification of students with unusual behavior, both those with learning problems and those who are gifted.

- **Clustering**: the grouping of related instances through the analysis of their resemblance (according to the values of their characteristics). Ranjan And Malik (2007) grouped students according to their educational history and socio-demographic characteristics. Clustering algorithms can be fuzzy, allowing an instance to belong to more than one algorithm, or disallow this. They can also start with no previous assumptions about what clusters exist, as is the case of K-Means (e.g., Steinbach, Karypis, Kumar, et al., 2000), or use an hypothesis as its starting point (e.g., R. Baker et al., 2010).

- **Relationship Mining**: the discovery of relationships between variables in large data sets and subsequent codification as rules that can be translated to other contexts. This technique was first introduced and applied by Agrawal et al. (1993). Hajizadeh and Ahmadzadeh (2014) use this technique to identify what are the effective factors on non re-taking a course by a student. A variation of this task is Sequential Pattern Mining, which focuses on codifying rules that map the connections between sequential events.

2.1.1 Academic Success

As of the writing of this document, the literature on academic success definition is scarce. Researchers have mainly focused their studies on predicting academic performance without classifying it as successful or not. Nghe, Janecek and Haddawy (2007) approach this by defining performance as a stratified concept, defining levels of performance and then predicting into which level a student would find himself in.

Understanding what are the crucial factors in a student's academic success is critical in determining what should be the predictor variables in such a model. Asif, Mercerin and Pathan (2015) analyze the comparative weight of sociodemographic factors in relation to the grades obtained both in pre-college and in the first year of college. Their analysis revealed that the grades were much more relevant for predicting performance, suggesting that sociodemographic data's main relevance lies in the prediction of first year results. This is supported by Mendes (2007), who shows that there is a very strong correlation between academic performance in high school and academic success in college. Another important aspect presented by Mendes is the increased predictive power of the number of courses completed by the student over the average grade he obtained in them.

Machado, Curado et al. (2006) however, appear to contradict the work of Asif, Mercerin and Pathan by showing a correlation between the economic and cultural status of students and their academic success, that appears to become stronger as their academic path progresses.

It is also important to note that the first year, due to its transitional nature, is also the one where students have the greatest probability of facing a lack of success or even attrition, and therefore one of the most important targets of analysis regarding academic success (Almeida & Cruz, 2010).

2.1.2 Data Mining Techniques in Educational Mining

Depending on the goal of the study different data mining techniques can be used to support the extraction of relevant knowledge. Taking into account the objectives of this study, we will focus our analysis on the classification and clustering techniques. In both cases there is no consensus regarding the most appropriate technique to use. Regarding classification, for example Ajay Kumar Pal and Saurabh Pal (2013) compared the performance of several algorithms and found the one with the best performance to be IB1, a Nearest Neighbour algorithm. Nghe, Janecek and Haddawy (2007), on the other hand, used Decision Trees (the j48 Weka implementation) and Bayesian Networks (the BayesNet Weka implementation) to predict student GPA at different points of their academic path, finding the Decision Trees consistently outperform the Bayesian Networks in this task. Chau and Phung (2013) compare several different algorithms in their work, showing Random Forest and Support Vector Machines to have the best performances, with Naive Bayesian Networks generaaly surpassing Decision Trees when dealing with imbalanced or discrete datasets.

Concerning classification problems, there also appears to be little consensus as to which set of predictors leads to the most accurate model. Oskuei and Askari (2014), for example, focus on the performance gains of using sex, parents' level of education and welfare, while Ajay Kumar Pal and Saurabh Pal (2013) utilize a broader range of attributes, taking into account aspects such as admission types and the locations of both the students' residence and the college. Meanwhile, Asif, Merceron and Pathan (2014) focused on the use of the grades students obtained in certain courses.

In the case of classification problems, the imbalaced nature of the dataset has deserved particular attention. For example, Chau and Phung (2013) use an approach that consists of oversampling the minority class and undersampling the majority class (this advantage of this approach over just oversampling the minority is that it maintains the performance gains while limiting the increase in size of the data set).

2.2 Conclusions

There is some variation in the literature as to what predictor variables are more relevant. This variation appears to be tied to the learning environment being studied, which seems to indicate that these models don't generalize well for other learning institutions. Furthermore, there is a lack of research on high-level prediction models in Universidade do Porto, with the closest being at the course level by Strecht et al. (2014). Therefore, a need for a prediction model on the degree level at Universidade do Porto can be identified.

The same lack of consistency in the literature regarding the predictor variables to use extends to the algorithms with better performance, which leads us to opt for performing our own empirical

analysis regarding how these two aspects perform over our dataset. On the other hand, the literature does present a very promising hybrid approach to resampling for dealing with the imbalanced nature of the dataset in the work of Chau and Phung (2013), though the point of balance will still need to be analyzed within the dataset we will be studying.

Chapter 3

Methodologies

The purpose of this project is to develop three distinct prediction models that address students academic performance. The first model concerns itself with predicting the academic success of a student on his/her first semester with the data available at the time of enrollment, while the second model uses the data available at the end of his/her first semester in order to predict his/her success at the end of the first academic year. For both of these models, success is defined as having completed at least 25 ECTS in the respective semester. This means that in a normal semester of 30 ECTS, the student gains approval in five-sixths of the courses. Both of these models use a cutoff point in the number of ECTS (in this case, 25), which is a continuous variable, in order to transform it into a binary one. This allows these models to perform as binomial classifiers. For the third model, however, a multinomial classifier was preferred, as this model focuses on predicting overall academic success of the student based on the data available at the end of the first year, and therefore knowing the degree of success was considered important. For measuring this performance value, the following formula was used:

$$PV = \frac{\sum_{(G*C)}}{\sum_{CE}} \tag{3.1}$$

with G representing the final grades the student obtained on his/her courses, C the number of ECTS those courses are worth and CE the total number of ECTS that s/he enrolled in. This allows us to take into account both how many courses s/he passed and the grades s/he obtained in them. The performance values calculated from the dataset were then submitted to a K-Means clustering algorithm, which allowed us to congregate them into five groups. By analyzing those groups, we were able to define the performance value ranges for five degrees of academic performance that were in accordance with the way the data shows that student's performance naturally groups itself, and use those degrees as our predictor classes. With regards to the information used to develop

these models, the attributes consisted of academic information, such as enrollment grades, national exams taken, amount of ECTS completed on a given semester and the average obtained, as well as demographic data, like a student's sex, his/her parents' education level and jobs, and whether he's a beneficiary of a scholarship.

For the sake of this project, four algorithms were chosen and used to build the three models of this project, namely C4.5 (through the j48 implementation from Weka), Random Forest, Naive Bayes and Support Vector Machine (through the SMO implementation from Weka). These algorithms were chosen due to their widespread use in data mining, as we can tell from the literature presented previously, and are described in the following sections along with the clustering algorithm used for the third model (K-Means).

3.1 Classification

Data Mining techniques for classification aim to predict the class or label of a data object, which is itself described by a set of attributes. Choosing these attributes is a crucial step in ensuring a good predictive performance.

3.1.1 C4.5

C4.5 is an algorithm for generating decision trees developed by Ross Quinlan (2014). Decision trees are a tree structure where the leaves represent class labels and the branches represent sets of features that lead to those leaves. According to Kotsiantis (2007), the algorithm is as follows:

1. Check for base cases
2. For each attribute x
   (a) Find the normalized information gain ratio from splitting on x
3. Let x_best be the attribute with highest normalized information gain
4. Create a decision node n that splits on x_best
5. Recur on sublists obtained from splitting on x_best, and add those as children of n

Information gain is defined as as being the non-symmetric measure of the difference between two probability distributions (Kullback & Leibler, 1951). When dealing with discrete probability distributions, it is given by

$$D_{KL}(P||Q) = \sum_i P(i) ln \frac{P(i)}{Q(i)} \tag{3.2}$$

with P representing the distribution of the observations and Q representing the model. The base cases for this algorithm are threefold:

- All the samples in the list belonging to the same class. This results in the creation of a leaf node representing the label for that class.
- No information gain is provided by the attributes available. This results in a decision node being created higher in the tree using the expected value of the class.
- Encountering an instance of a previously-unseen class. This also results in a decision node being created higher in the tree using the expected value of the class.

3.1.2 Naive Bayes

Naive Bayes classifiers are a family of simple probabilistic classifiers based on applying Bayes' theorem with strong (naive) independence assumptions between the features. Bayes' theorem describes the probability of an event, based on conditions that might be relevant to the event, as given by this formula:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{3.3}$$

Given an instance to be classified, represented by a vector $\mathbf{x} = (x_1, \ldots, x_n)$ representing some n features (independent variables), Naive Bayes assigns to this instance probabilities $p(C_k|x_1, \ldots, x_n)$ for each of k possible outcomes or classes (Murty & Devi, 2011).

Naive Bayes assumes that each feature F_i is independent of every other feature F_j, for $I \neq J$, hence the it being categorized as "naive". With this assumption, we can express the conditional distribution over class C as:

$$p(C_k|x_1, \ldots, x_n) = \frac{1}{Z} p(C_k) \prod_{i=1}^{n} p(x_i|C_k) \tag{3.4}$$

where $Z = p(\mathbf{x})$ is a scaling factor dependent only on $x_1, \ldots, x_n$, that is, a constant if the values of the feature variables are known.

3.1.3 Support Vector Machine

A Support Vector Machine (SVM) is a supervised learning model that, given a set of examples belonging to one of two categories, constructs a hyperplane in a high- or infinite-dimensional space, which can be used to assign future examples into one of those categories. A good separation is achieved by the hyperplane that has the largest distance to the nearest training-data point of any class (so-called functional margin), since in general the larger the margin the lower the generalization error of the classifier (Cortes & Vapnik, 1995).

Given some training data $\mathscr{D}$, a set of n points of the form

$$\mathscr{D} = \{(\mathbf{x}_i, y_i) | \mathbf{x}_i \in \mathbb{R}^p, y_i \in \{-1, 1\}\}_{i=1}^n \tag{3.5}$$

where the y_i is either 1 or -1, indicating the class to which the point $\mathbf{x}_i$ belongs. Each $\mathbf{x}_i$ is a p-dimensional real vector. We want to find the maximum-margin hyperplane that divides the points having $y_i = 1$ from those having $y_i = -1$. Any hyperplane can be written as the set of points $\mathbf{x}$ satisfying $\mathbf{w} \cdot \mathbf{x} - b = 0$, where $\cdot$ denotes the dot product and $\mathbf{w}$ the (not necessarily normalized) normal vector to the hyperplane. The parameter $\frac{b}{|\mathbf{w}|}$ determines the offset of the hyperplane from the origin along the normal vector $\mathbf{w}$.

If the training data are linearly separable, we can select two hyperplanes in a way that they separate the data and there are no points between them, and then try to maximize their distance. The region bounded by them is called "the margin". These hyperplanes can be described by the equations

$$\mathbf{w} \cdot \mathbf{x} - b = 1 \tag{3.6}$$

and

$$\mathbf{w} \cdot \mathbf{x} - b = -1 \tag{3.7}$$

By using geometry, we find the distance between these two hyperplanes is given by $\frac{2}{|\mathbf{w}|}$, so we want to minimize $|\mathbf{w}|$ in order to maximize that distance. As we also have to prevent data points from falling into the margin, we add the following constraint: for each i either

$$\mathbf{w} \cdot \mathbf{x}_i - b \geq 1 \qquad \text{for } \mathbf{x}_i \text{ of the first class} \tag{3.8}$$

or

$$\mathbf{w} \cdot \mathbf{x}_i - b \leq -1 \qquad \text{for } \mathbf{x}_i \text{ of the second} \tag{3.9}$$

This can be rewritten as:

$$y_i(\mathbf{w} \cdot \mathbf{x}_i - b) \geq 1, \quad \text{for all } 1 \leq i \leq n. \qquad (1) \tag{3.10}$$

This can be used to express the following optimization problem:
Minimize (in $\mathbf{w}, b$)
$|\mathbf{w}|$ subject to (for any $i = 1, \ldots, n$)
$y_i(\mathbf{w} \cdot \mathbf{x_i} - b) \geq 1$.
This optimization problem can be expressed in the dual form as

$$\max_{\alpha} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} y_i y_j K(x_i, x_j) \alpha_i \alpha_j$$

(3.11)

Where x_i is the input vector, $y_i \in \{-1;+1\}$ a binary label corresponding to it and the variables α_i are Lagrange multipliers.

Sequential Minimal Optimization is an algorithm that solves the above problem by dividing it into a series of smaller sub-problems (Platt, 1998). The algorithms consists of the following steps:

1. Find a Lagrange multiplier α_1 that violates the Karush–Kuhn–Tucker (KKT) conditions for the optimization problem. The KKT conditions are first order necessary conditions for a solution in nonlinear programming to be optimal
2. Pick a second multiplier α_2 and optimize the pair (α_1,α_2)
3. Repeat the previous steps until convergence occurs

The problem is solved when all Lagrange multipliers satisfy the KKT conditions.

3.2 Clustering

Data Mining techniques for clustering focus on grouping a set of objects into groups (called clusters) such that objects in one group are more similar (as defined by their attributes) to one another than to objects of other groups. While there are several different clustering algorithms, one of the most popular and easier to understand is K-Means.

3.2.1 K-Means

K-Means is an algorithm that aims to partition n instances into k clusters, with each instance belonging to the nearest cluster.

David MacKay (2003) defines the K-Means algorithm as follows:

1. Initialization: Set K means (m^k) to random values
2. Assignment step: Each data point n is assigned to the nearest mean. We denote our guess for the cluster k^n that the point x^n belongs to by k'^n.

$$k'^n = min_k d(m^k, x^n)$$

(3.12)

An alternative, equivalent representation of this assignment of points to clusters is given by 'responsibilities', which are indicator variables r_k^n. In the assignment step, we set r_k^n to one if mean k is the closest mean to data point x^n; otherwise r_k^n is zero.

3. Update step: The means are adjusted to match the sample means of the data points they are responsible for.

$$m^k = \frac{\sum_n r_k^n x^n}{R^k} \tag{3.13}$$

Where R^k is the total responsibility of mean k,

$$R^k = \sum_n r_k^n \tag{3.14}$$

4. Repeat the assignment step and the update step until the assignments do not change.

3.3 Data Preparation and Performance Evaluation

As an essential step in a data mining process, data preparation involves a series of tasks that aim to improve the quality of the data by removing or correcting corrupt, inaccurate or missing records, as well as by transforming the data itself, for example through the removal of redundant attributes or by resampling to balance the dataset. The specific data preparation techniques utilized in this project are explored in the next chapter.

The algorithms themselves were encapsulated in a process that applies x-fold validation with 10-folds, meaning that the data was divided into 10 blocks, the model was trained 9 of the blocks and evaluated with the other one. The process was repeated 10 times, once for each of the different blocks. In the end, the average performance was used. This reduces the impact of the selection of data for training and test sets has on the performance of the model.

For the performance itself, three different measures were used in the first two models: Accuracy, Area Under Curve (AUC) and Specificity. Since in the first year priority is given to identifying cases of unsuccessful students rather than successful ones, as these are the ones the institutions need to support, Specificity is more relevant in those cases than Sensitivity, which prioritizes the identification of successful examples. For the third model, Accuracy, Sensitivity (also called Recall) and Precision. Next we present a brief description of the evaluation metrics used:

- Accuracy: corresponds to the number of correct predictions divided by the total number of predictions. This measure can be misleading on unbalanced datasets, but when the dataset is balanced, gives a good indication of how well the model can identify the classes;

- Specificity: corresponds to the number of negatives correctly identified as such divided by the total number of negatives (both those identified as such and those identified as positives). Essentially, this corresponds to a measure of the model's avoidance of false positives;
- Sensitivity: corresponds to the number of positives correctly identified as such divided by the total number of positives (both those identified as such and those identified as negatives);
- Precision: corresponds to the number of positives correctly identified as such divided by the total number of instances identifies as positives (both true and false positives);
- Area Under Curve: the ROC (Receiver Operating Characteristic) curve results from plotting the Sensitivity (give by the number of positives correctly identified as such divided by the total number of positives) against 1 - Specificity. A random classifier should have an AUC of 0.5, meaning it correctly classifies half the instances, while a perfect classifier would have an AUC of 1.

For the development of this project RapidMiner was the chosen platform. This platform was chosen due to the following characteristics:

- the drag-and-drop nature of the framework, which greatly speeds up development by providing ready-made processes;
- the integration of third party plugins, namely from Weka, which provides algorithms that are known to be efficient;
- the ability to write processes in R, which allows the extension of the framework to deal with cases not covered by the available processes.

It is important to note that Rapidminer uses a cut-off point of 0.5 to identify the positive cases, used for example to estimate the Accuracy.

The framework for the algorithms used can be seen in figure 3.1. There we can see that in all instances, the data went through a block of data preparation processes, with slight variations (the randomForest algorithm didn't involve going through feature selection).

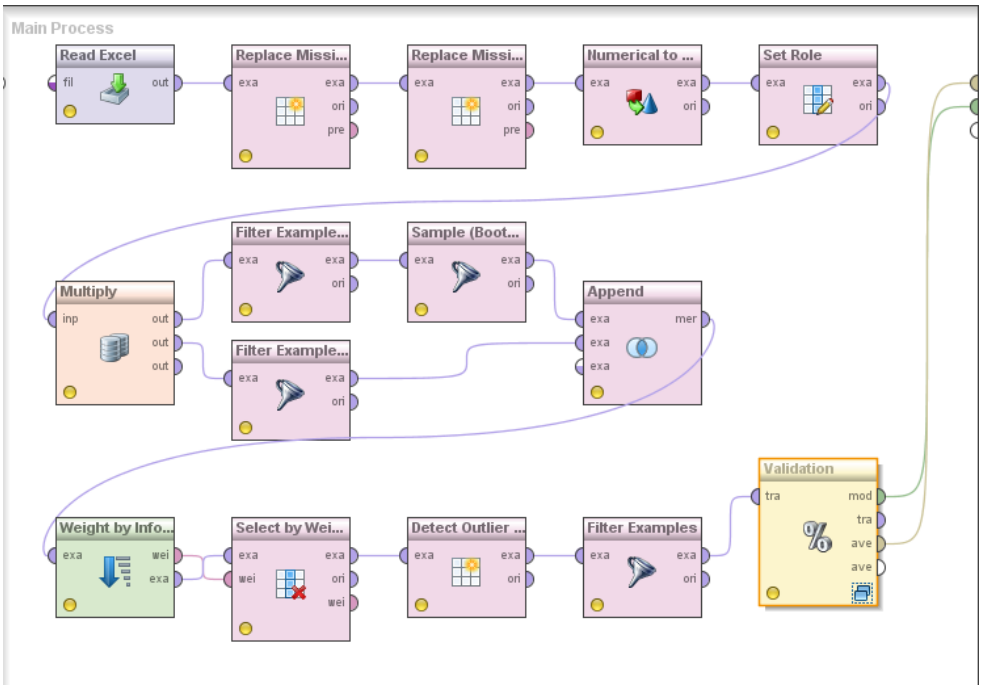

Figure 3.1: Framework

Chapter 4

Data Exploration and Preparation

After studying the domain, the next step in a Data Mining approach usually involves performing an exploratory analysis of the data available. Often, simple statistical analysis of the attributes in a dataset and their relationship to the predictive classes yields a lot of relevant information even before Data Mining algorithms are used.

Regarding the dataset used for this thesis, it was made available by FEUP from their academic information on students who enrolled between 2003 and 2007, accompanying those students all the way to either the conclusion of their degrees or the year 2015, whichever came first. The model for the dataset can be seen in figure 4.1.

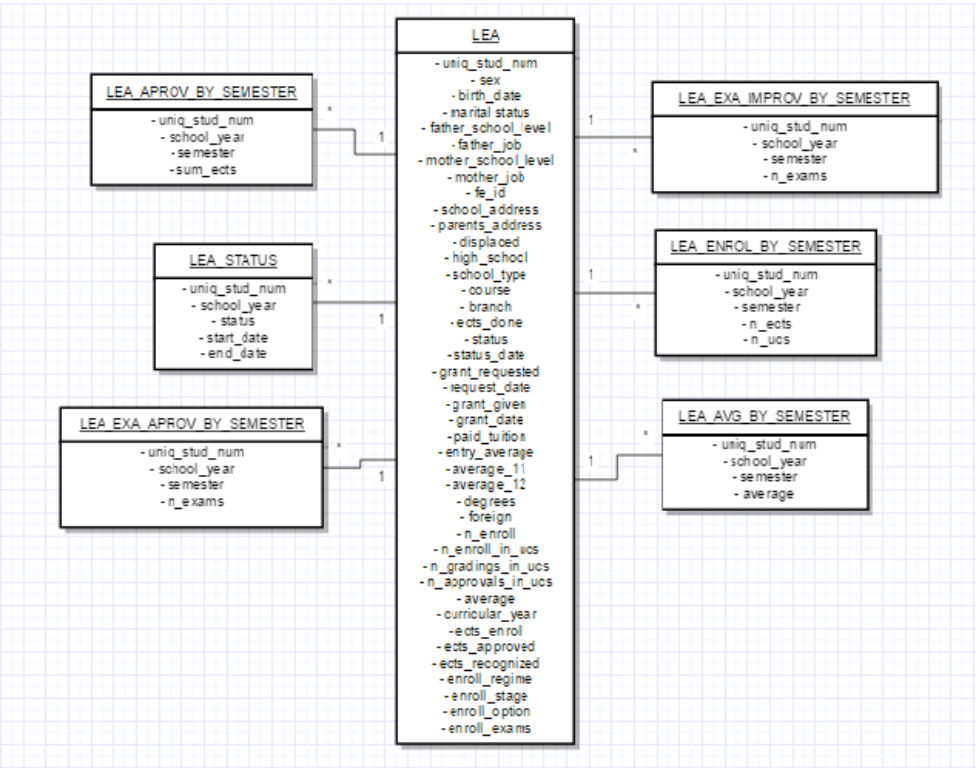

Figure 4.1: Data Model

From the above model it is apparent that information regarding student involvement with academic life is not present, which excludes it's inclusion in the models developed in the present work. Another aspect that is missing, perhaps more importantly, is information regarding which courses the student has completed, which eliminates the possibility of identifying those that merit more attention by the decision makers.

By analyzing the model, some aspects stand out as providing valuable avenues of study, namely:

- sociodemographic information about the student and his/her socio-cultural status (provided by the student's sex, his/her parents' jobs and educational levels, the type of school s/he attended in high school, whether s/he is displaced or not and information regarding grant requests and payment of tuition);
- academic information, be it with regards to the student's past performance (such as high school and enrollment average grades, the enrollment exams performed, average grades and number of ECTS completed in each semester of the first year in college), administrative data (such as enrollment year and the degree the student enrolled in) or both performance and motivation (seen in the enrollment stage and option).

An important thing to note in this model is how the information regarding whether tuition was paid, a grant was requested/given or if a student is displaced is not related to the school year, meaning variation in this information along a student's academic life is not accounted for, reducing the usefulness of this data, especially as it makes it impossible to know whether the information is applicable to the student's first year, which is the focus of two of the developed models. This is not the case of the information regarding the student's status, where the information is connected to a specific semester.

Variables that identify the student himself rather identifying his/her characteristics, such as his/her unique student number, his/her ID number at FEUP, and his/her birth date are not used in these models, as well as any information that would not be available at the time where the prediction is supposed to take part (such as information on the student's diplomas).

4.1 Population analysis

The first step in comprehending the dataset is categorizing the population it entails. For these models, we took into account only students who had enrolled in the regular contingent, ignoring the special enrollment cases, while also taking into account only the degrees that correspond to Licentiates or Masters. These filters also eliminated any foreign students from our population.

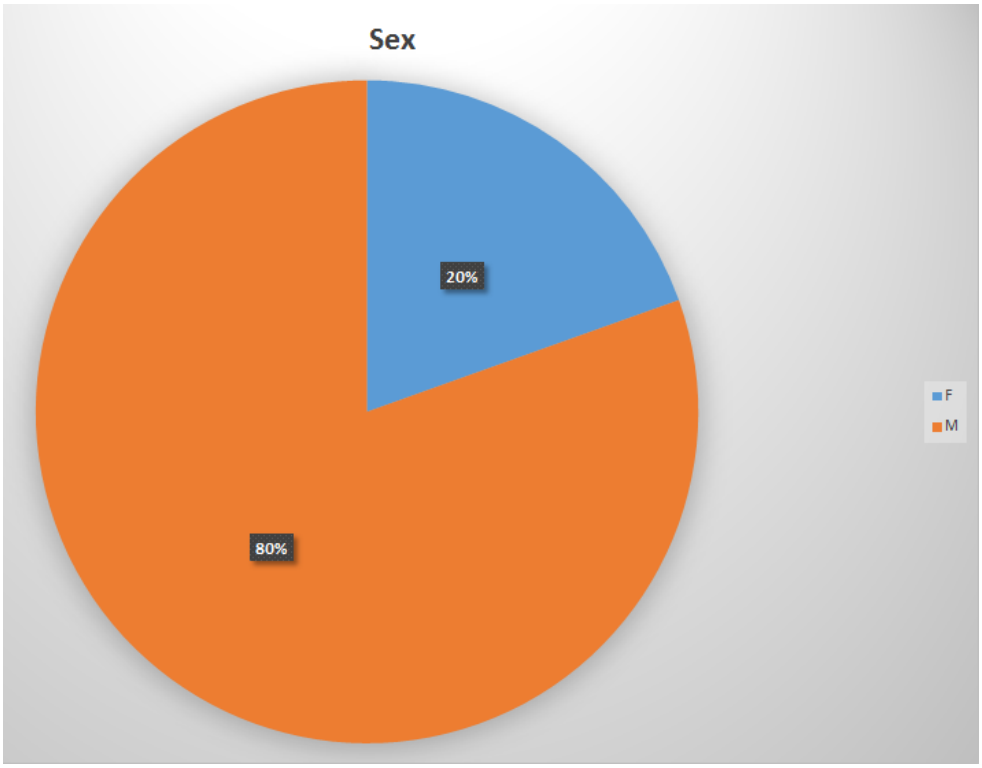

Figure 4.2: Ratio of Males/Females

In figure 4.2 we see that the population for the first two models is predominantly male, while figure 4.3 shows us that most students come from public schools.

In figure 4.4 we can see that there's a reasonable representation of a variety of degrees, meaning the dataset isn't overly unbalanced towards a single degree.

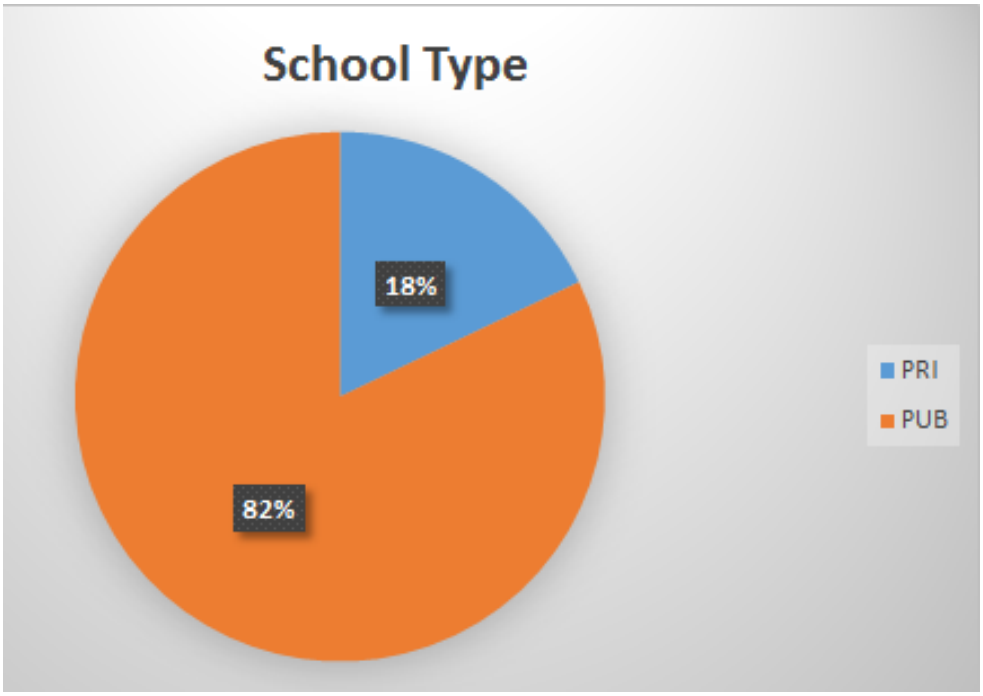

Figure 4.3: Ratio of School Type

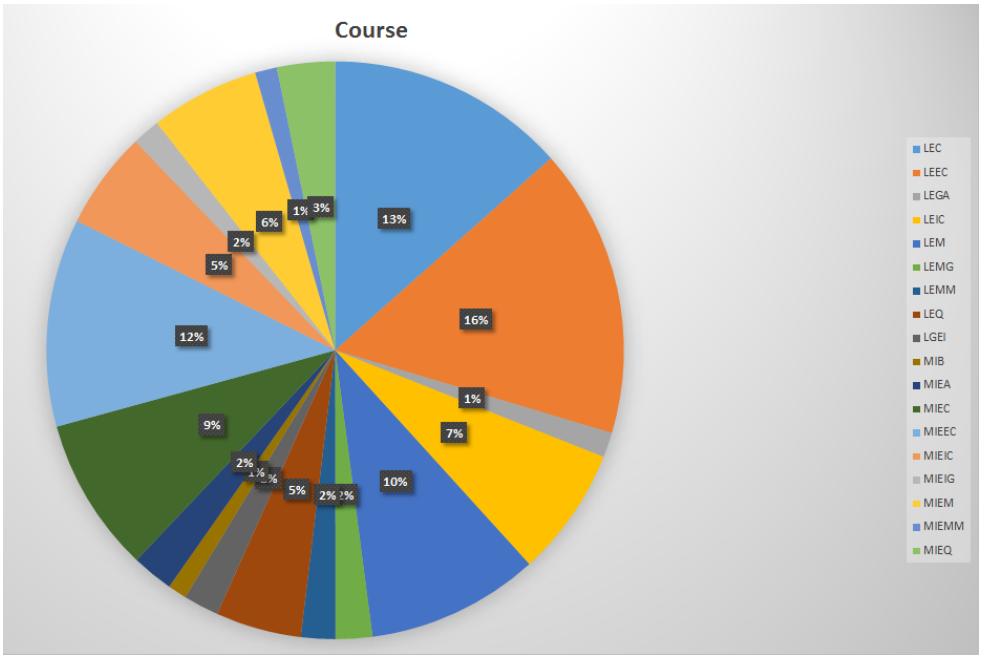

Figure 4.4: Ratio of Degrees

Figure 4.5 shows that the mean enrollment average grade is of 14.86, with a standard deviation of 1.63. It is important to note that the extremes of 10 and 20 having almost no instances. High School average grades, on the other hand, are generally a little higher, having a mean value of 15.53 and a standard deviation of 1.60, as seen in figure 4.6.

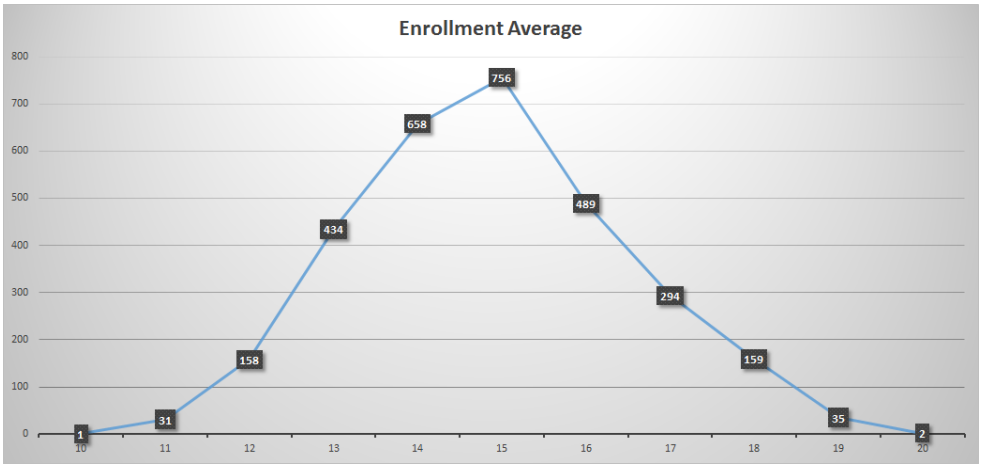

Figure 4.5: Distribution of Enrollment Average Grade

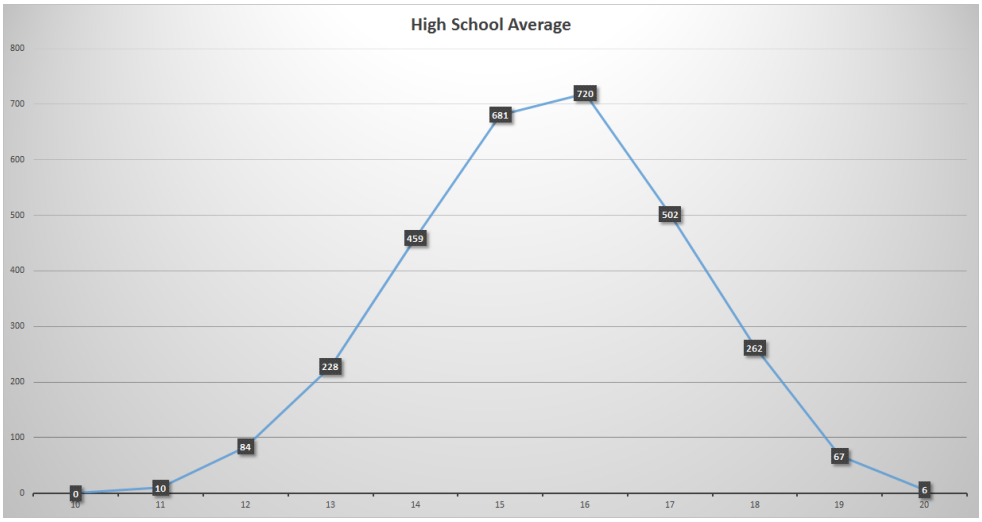

Figure 4.6: Distribution of High School Average Grades

We can see from figure 4.7 that most of the population enrolled in the first stage, and also that we have no instances of third stage students.

As for the enrollment option, we can see in figure 4.8 that the overwhelming majority of students were able to enroll in their first or second options.

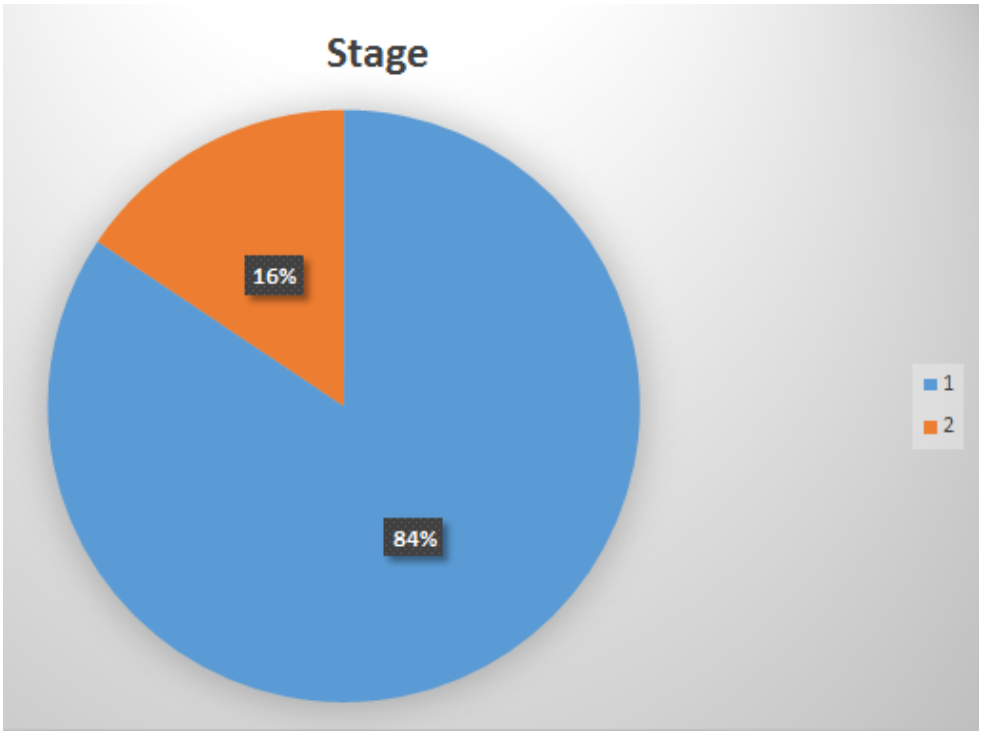

Figure 4.7: Ratio of Enrollment Stage

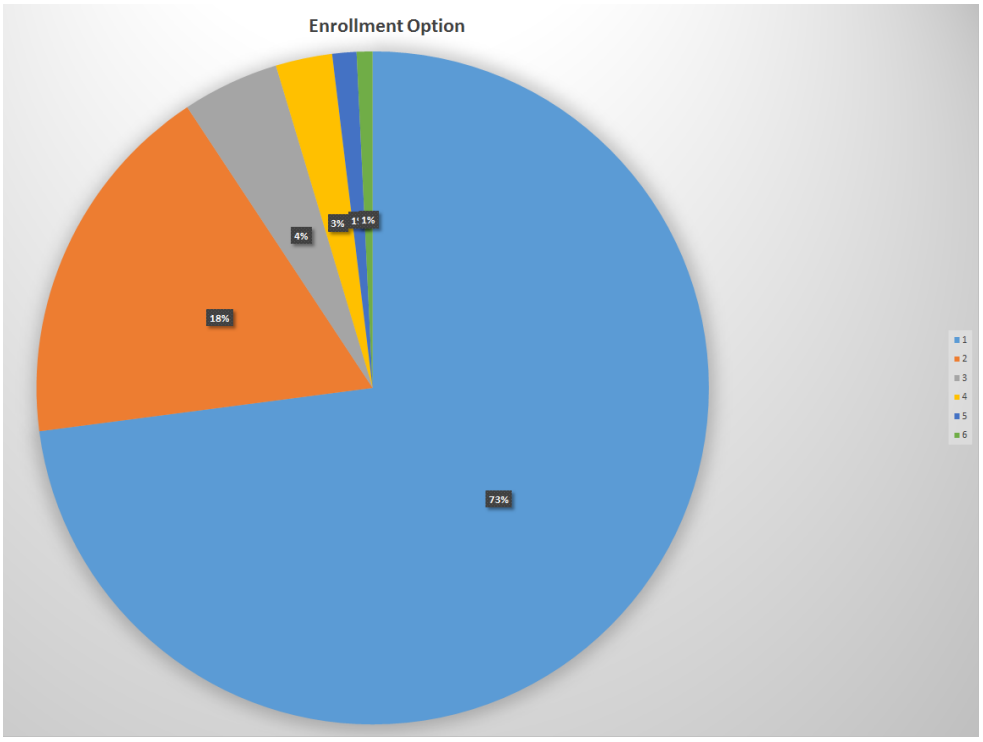

Figure 4.8: Ratio of Enrollment Option

Lastly, in figure 4.9 we can see that all five enrollment years are present in the dataset in a balanced way, which should keep the algorithms from overfitting to a specific year's characteristics.

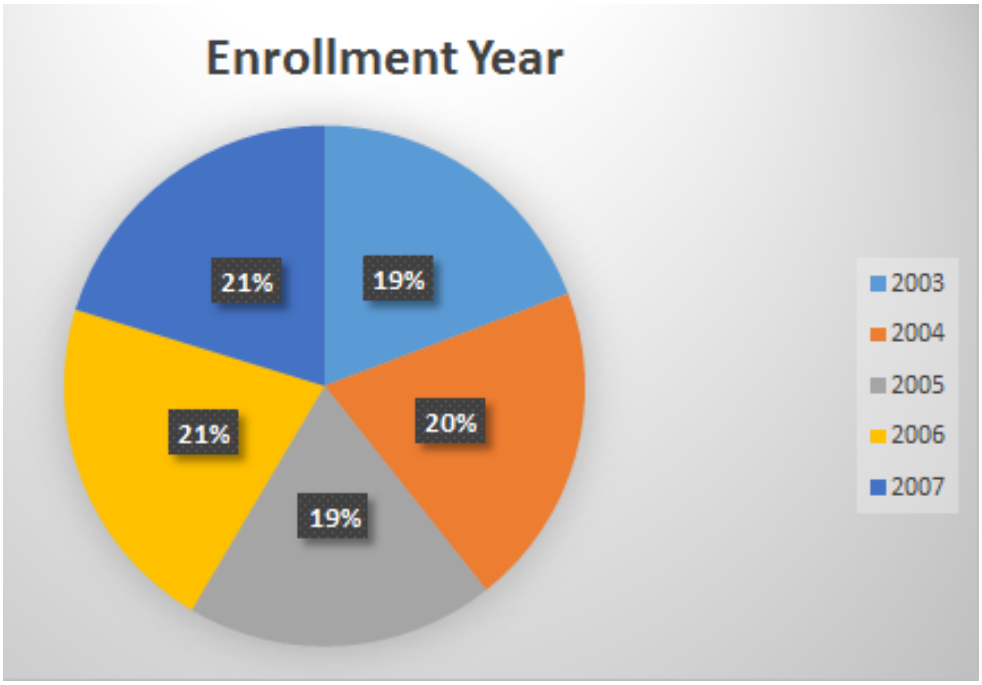

Figure 4.9: Ratio of Enrollment Year

We can therefore conclude that the population is composed predominantly of male students, from public schools, who enrolled in the first stage on their first or second options, with an average grade in the range of 13-17.

Another very important realization in this stage of the analysis was the discovery of missing values regarding parents' education level and job, displacement and 11th year average grade. While the parent's information was completed with a default value of "unknown" denoting a lack of information, the other attributes were eliminated from the dataset (the addresses and displacement due to the lack of reliability caused by the information only being collected at the exact time of enrollment, and the 11th year average due to the existence of good information regarding the 12th year average grade, also known as high school average grade).

4.2 Success on first semester

Having analyzed the population, it is now important to understand how their characteristics relate to the prediction class. In order to study those relationships, for every model developed these characteristics will be mapped against the success measure.

As mentioned previously, success for this model is defined as completing at least 25 ECTS.

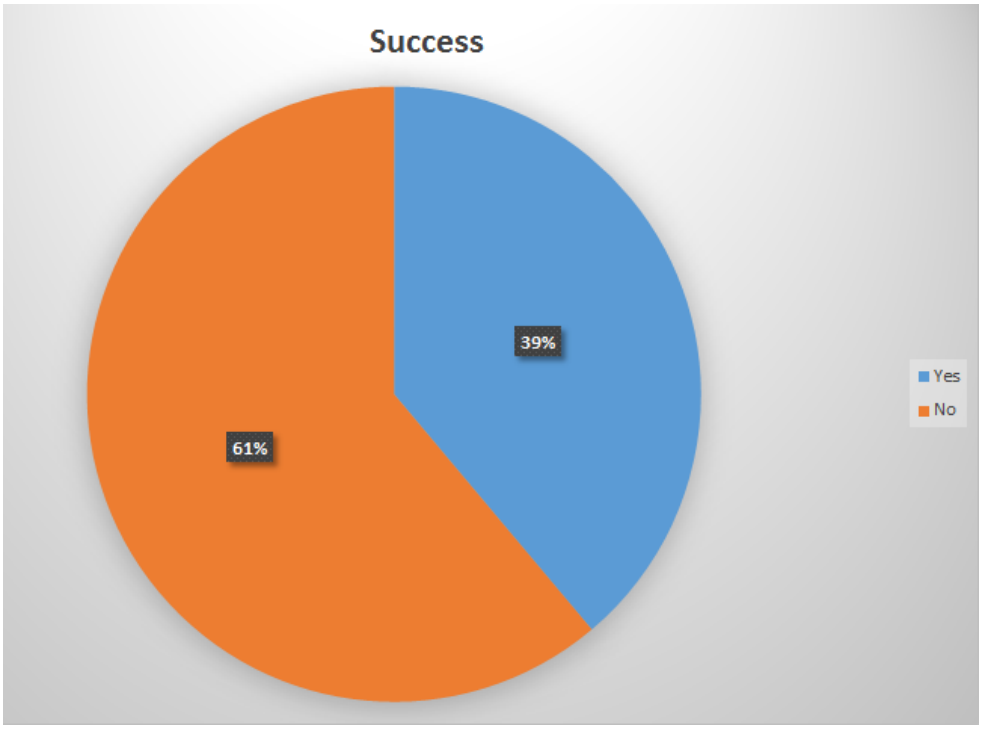

Figure 4.10: Ratio of Success on First Semester

The ratio of success on the first semester shows some unbalance, with more students being unsuccessful as seen in figure 4.10. This might create a bias in the predictive algorithms, which should be eliminated through resampling techniques.

Regarding the student's sex and it's relationship with success, we can see that males have a bias towards lack of success (see figure 4.11). Nonetheless, considering the ratio of males and females, and that females also show a (admittedly smaller) bias for not succeeding, this attribute's impact might not be very relevant.

Figure 4.12 shows that school type doesn't appear to have any effect on success on the first semester, which seems to contradict the idea that private schools prepare their students better for college. The lack of impact of this attribute also suggests that feature selection might be an important step in the Data Preparation stage.

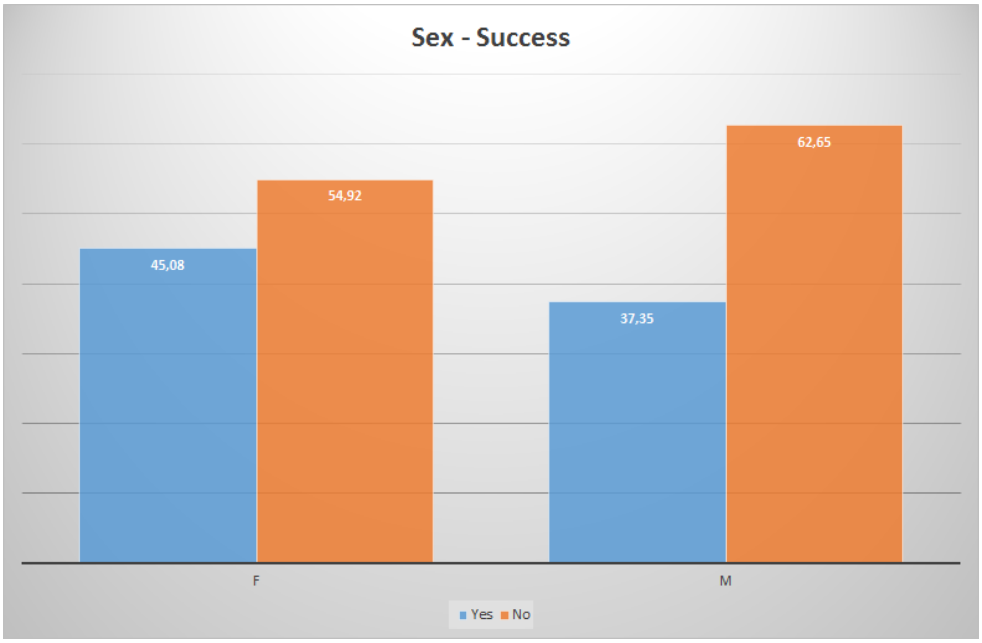

Figure 4.11: Ratio of Success on First Semester by Sex

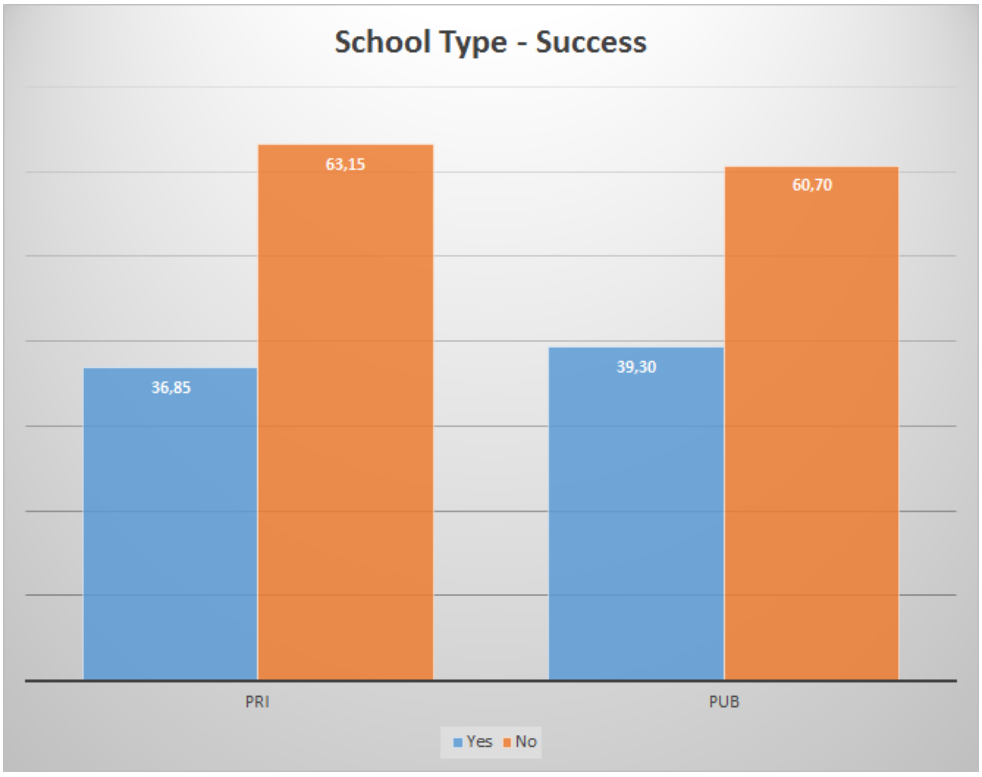

Figure 4.12: Ratio of Success on First Semester by School Type

In figure 4.13 we can see that while some degrees, most notably LEMG, MIEEC, MIEQ and LGEI, show a large bias regarding the success of the students enrolled in them, and others show a similar bias for the lack of success, such as MIEQ, MIEEC and LEMG, there are some, like LEC, MIB and MIEM, that appear to have no impact at all.

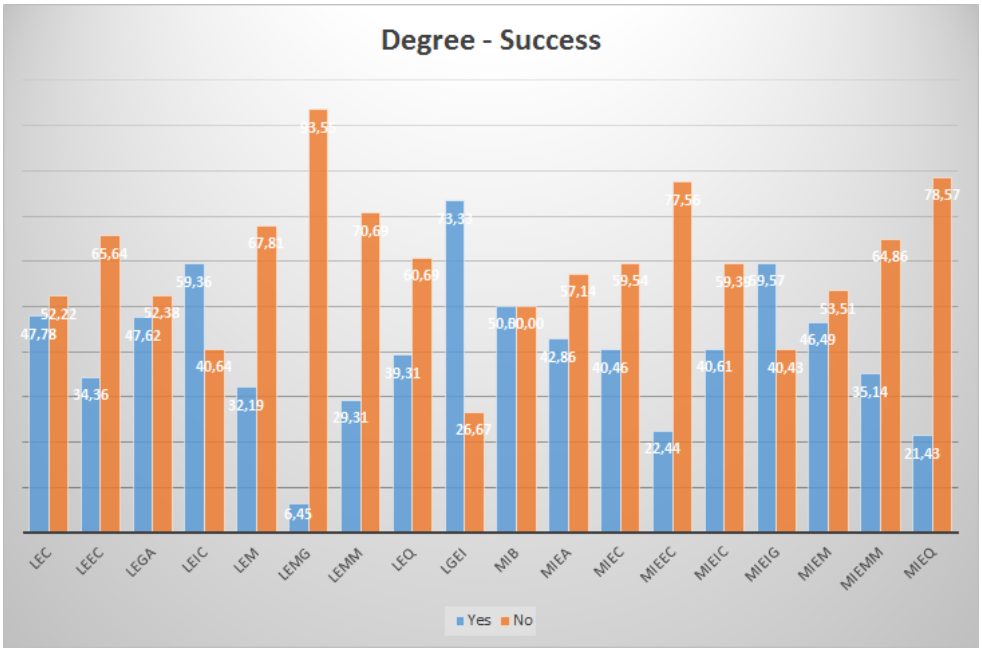

Figure 4.13: Ratio of Success on First Semester by Degree

The High School grades, as would be expected, have a very large correlation to the success on the first semester, both the enrollment average grade (figure 4.14) and the High School average grade (figure 4.15).

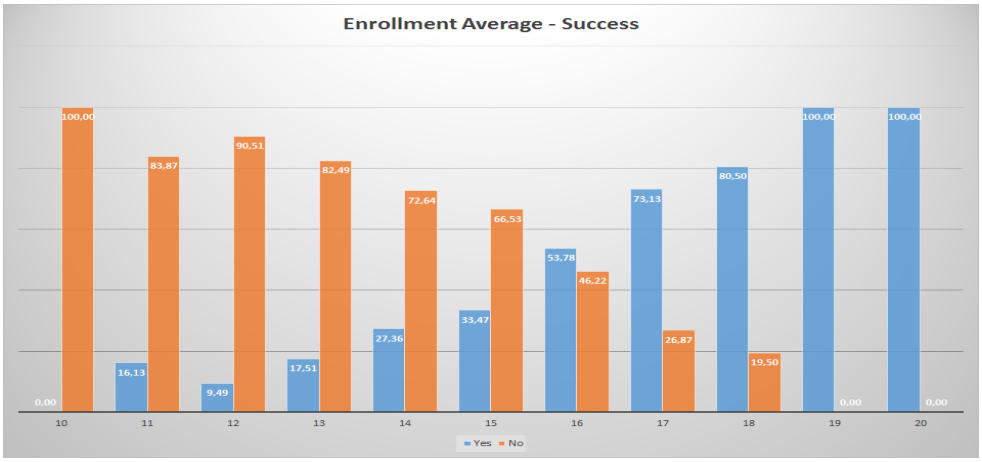

Figure 4.14: Ratio of Success on First Semester by Enrollment Average

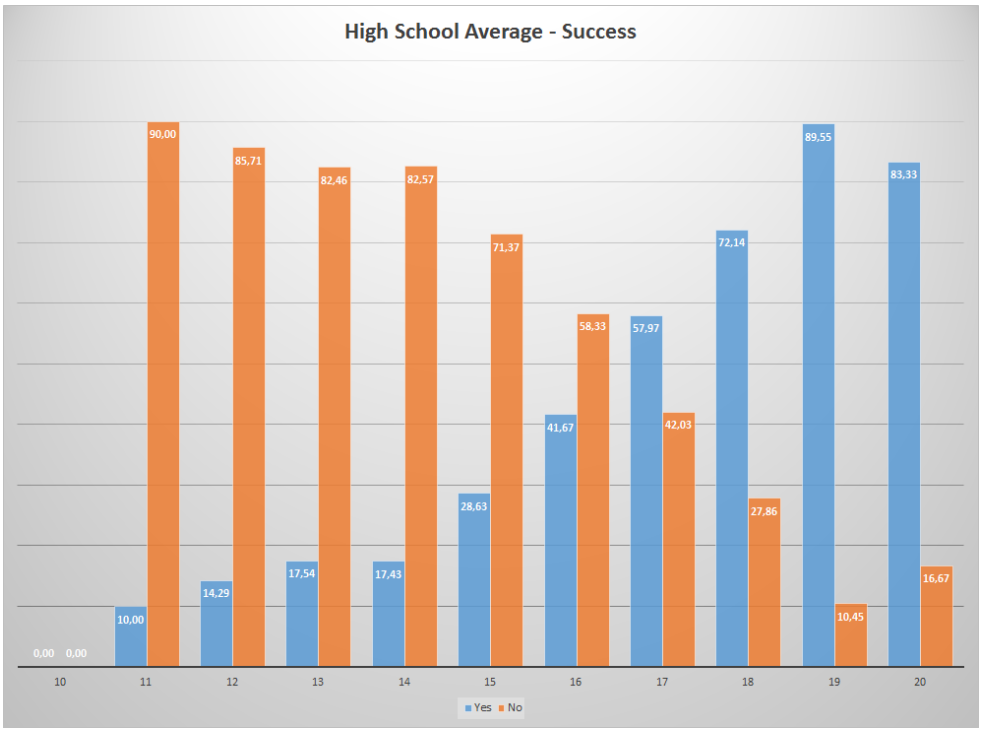

Figure 4.15: Ratio of Success on First Semester by High School Average

Enrollment Stage and Enrollment Option also show a strong correlation to academic success, with success being lower the later the stage or the lower the option, which is likely to be related to the motivation of the student for completing that degree (see figures 4.16 and 4.17).

Lastly, figure 4.18 shows that there doesn't seem to be any relationship between the year the student enrolled in and his/her success on the first semester. Once again, the existence of attributes with apparently little correlation to the prediction class points towards feature selection being a good way to enhance prediction performance.

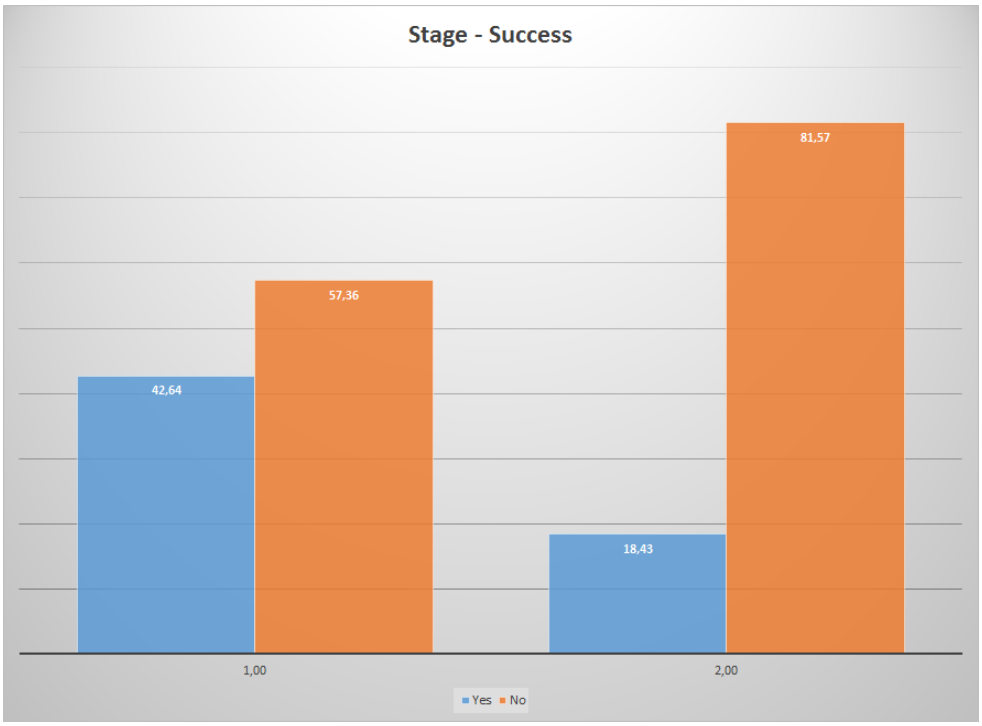

Figure 4.16: Ratio of Success on First Semester by Enrollment Stage

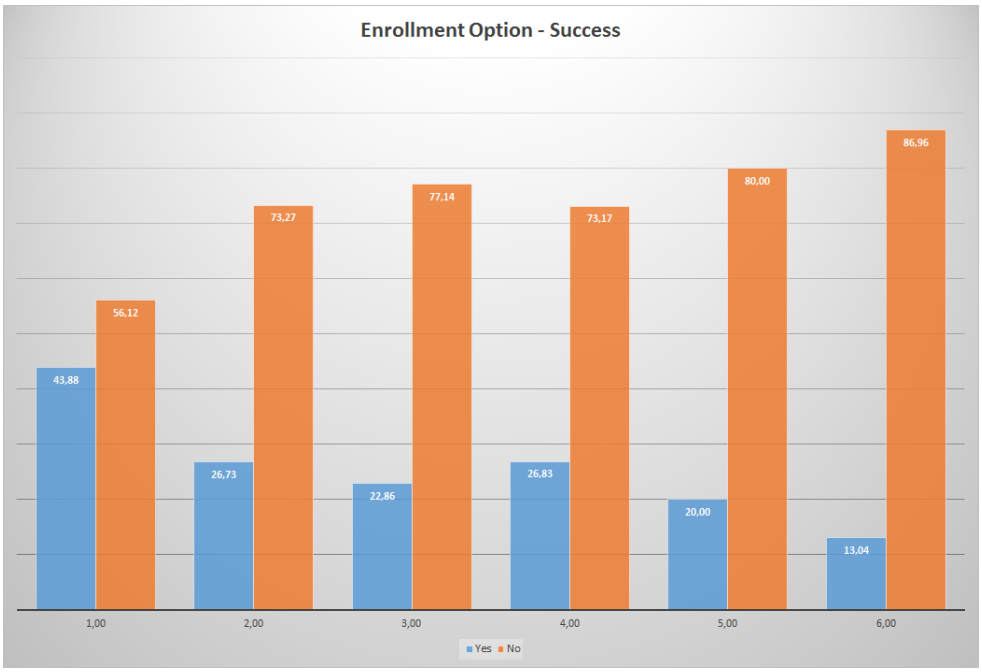

Figure 4.17: Ratio of Success on First Semester by Enrollment Option

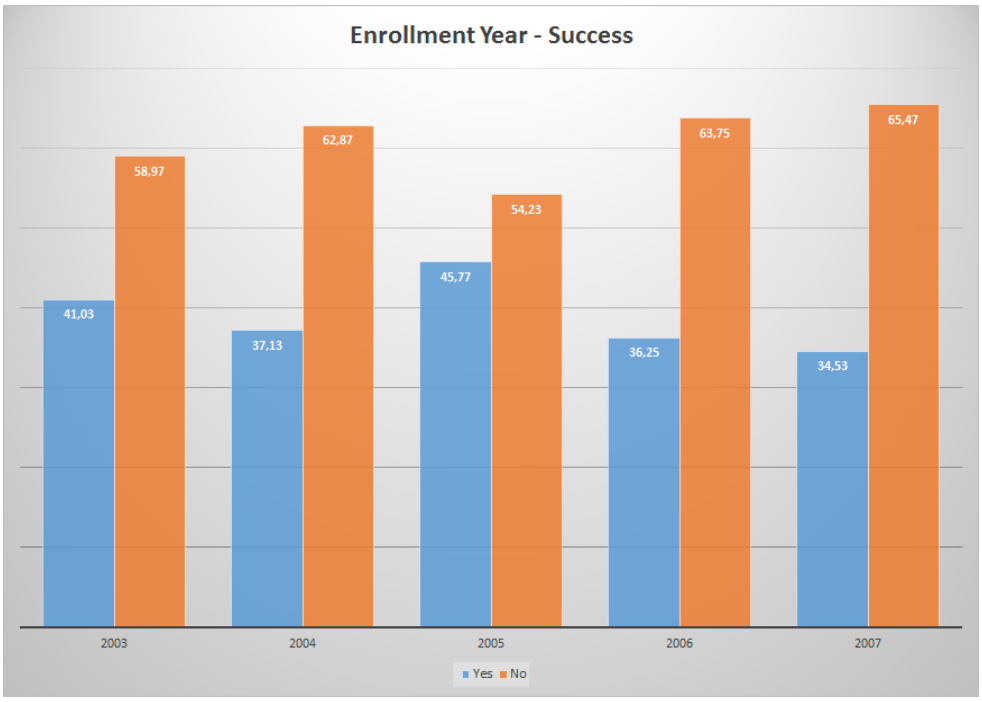

Figure 4.18: Ratio of Success on First Semester by Enrollment Year

From this analysis, we can conclude that either the enrollment average grade or the high school average grade should have a very high predictive impact, with the impact of the enrollment stage and the enrollment option being more limited due to the low amount of instances where these values differ from the predominant values. We also identify resampling and feature selection as important processes to apply to this dataset.

4.3 Success on second semester

As mentioned previously, success on the second semester is defined as completing at least 25 ECTS.

Similarly to the first semester, the ratio of success on the second semester is also unbalanced, favoring lack of it (figure 4.19). Once again, this suggests that resampling should be used.

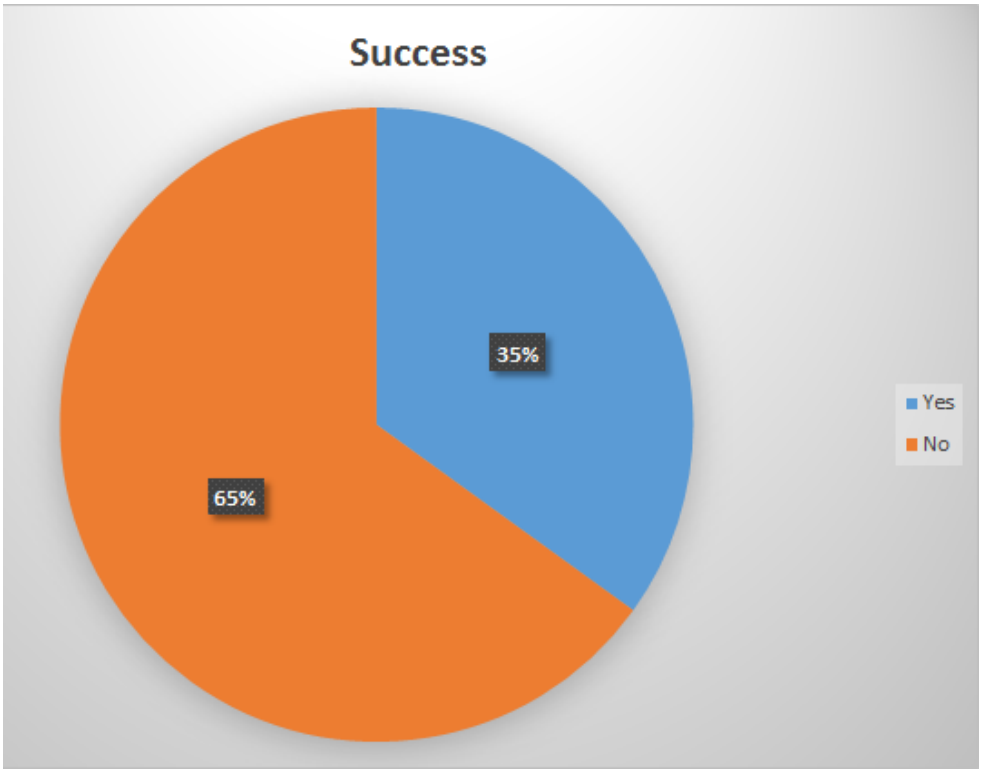

Figure 4.19: Ratio of Success on Second Semester

Once again, the student's sex and the type of school attended in high school do not appear to be highly correlated to the academic success in the second semester (figures 4.20 and 4.21). Like for the first semester, this suggests the use of feature selection.

More so than what happened in the first semester, there appears to be an imbalance regarding the academic success of the students in the various degrees (see figure 4.22). While some do still seem to have very little weight, such as LEC, LEIC and LEM, it appears this attribute is even more likely to have an impact on the second semester.

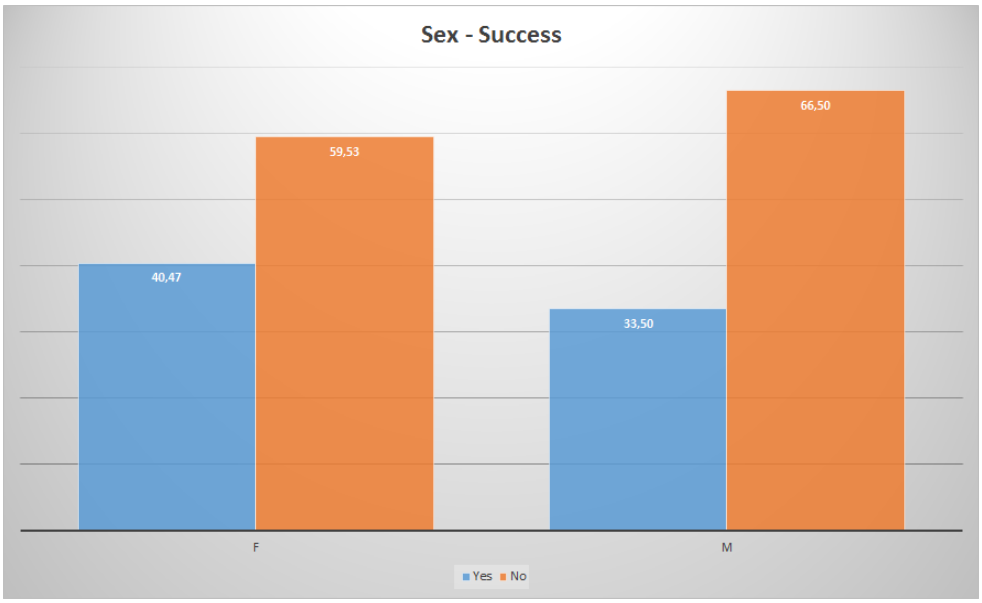

Figure 4.20: Ratio of Success on Second Semester by Sex

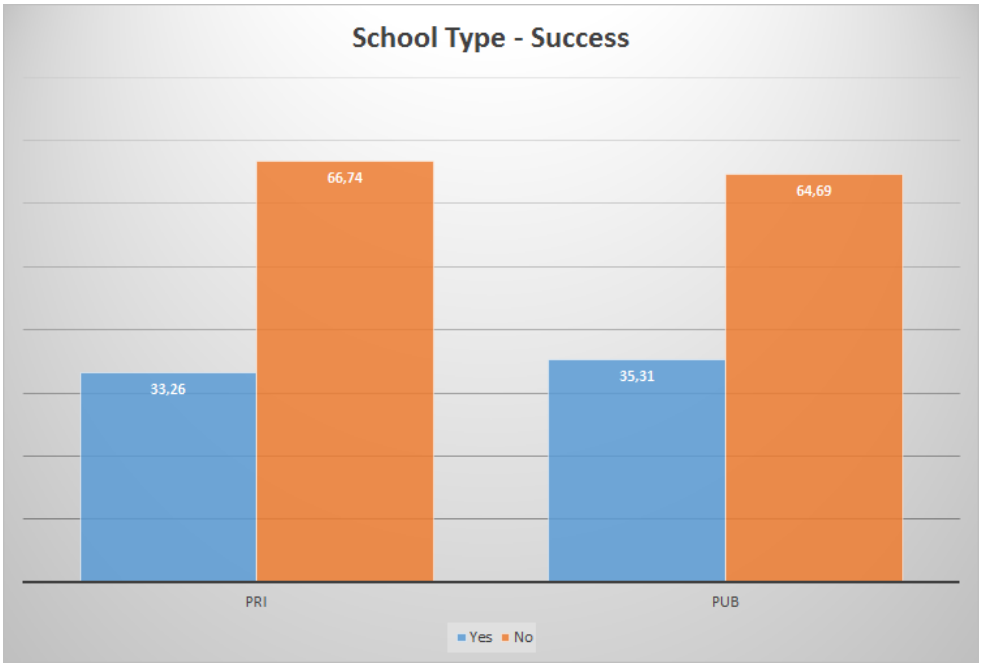

Figure 4.21: Ratio of Success on Second Semester by School Type

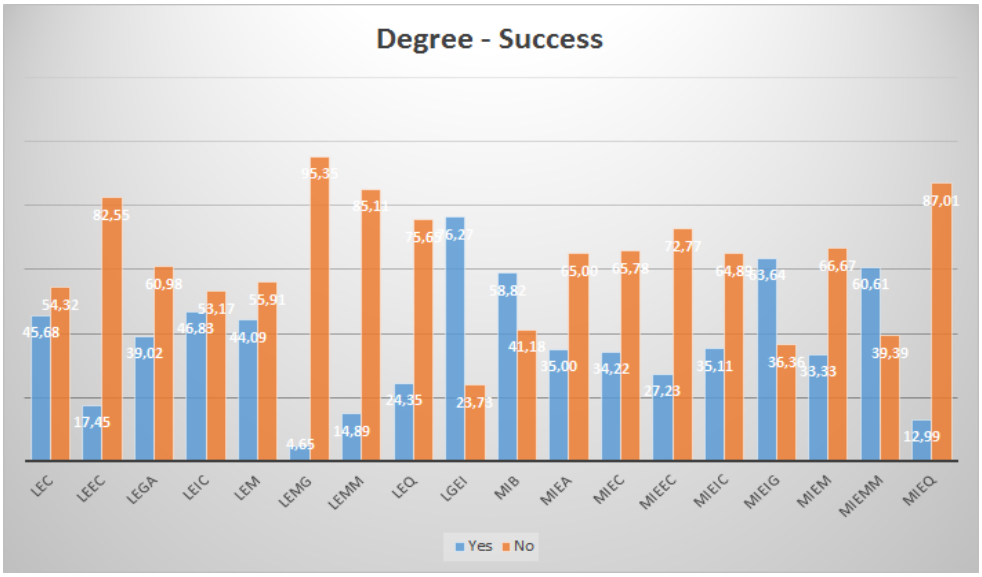

Figure 4.22: Ratio of Success on Second Semester by Degree

The predictive impact of the enrollment average grade (figure 4.23) and the high school average grade (figure 4.24) seems similar to what happened on the first semester. However, considering that the data related with this success on the first semester is available for this model, these attributes might end up not having as much impact.

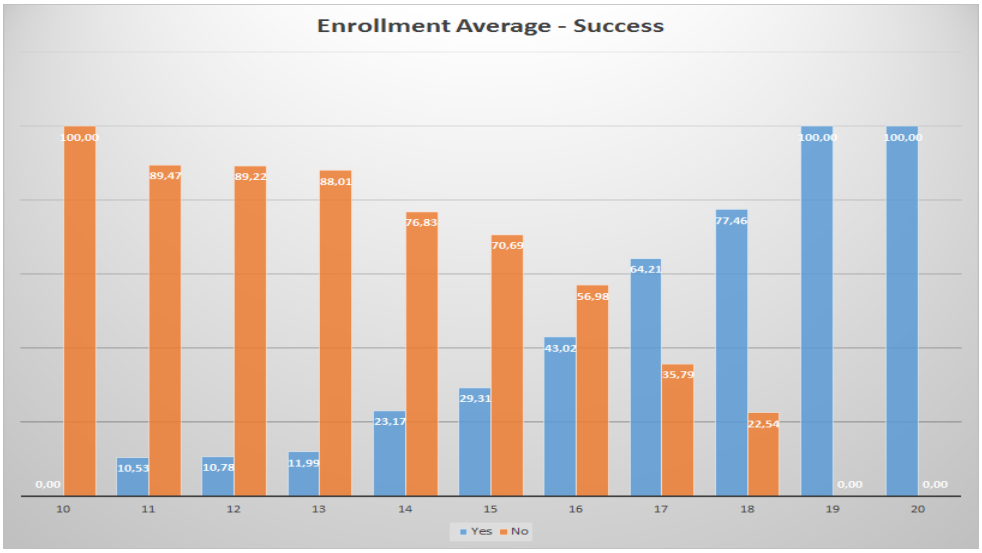

Figure 4.23: Ratio of Success on Second Semester by Enrollment Average

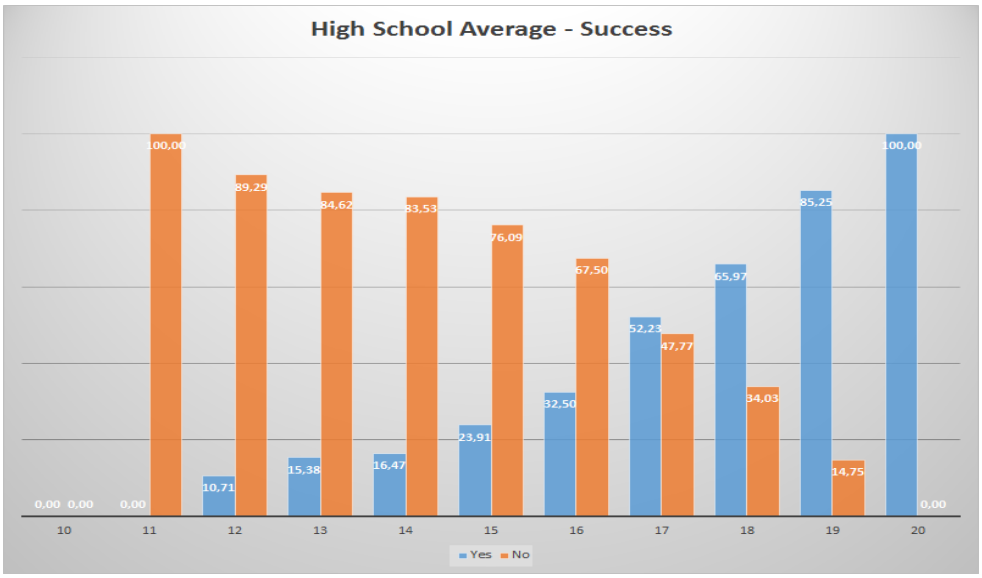

Figure 4.24: Ratio of Success on Second Semester by High School Average

As we can observe on figures 4.25 and 4.26, there seems to be some correlation between the enrollment stage and option and the academic success of the student, though this correlation seems weaker for the second semester than for the first. An important thing to note here is how the 6th enrollment option has a bias towards success, contrary to what would be expected and was seen in the first semester.

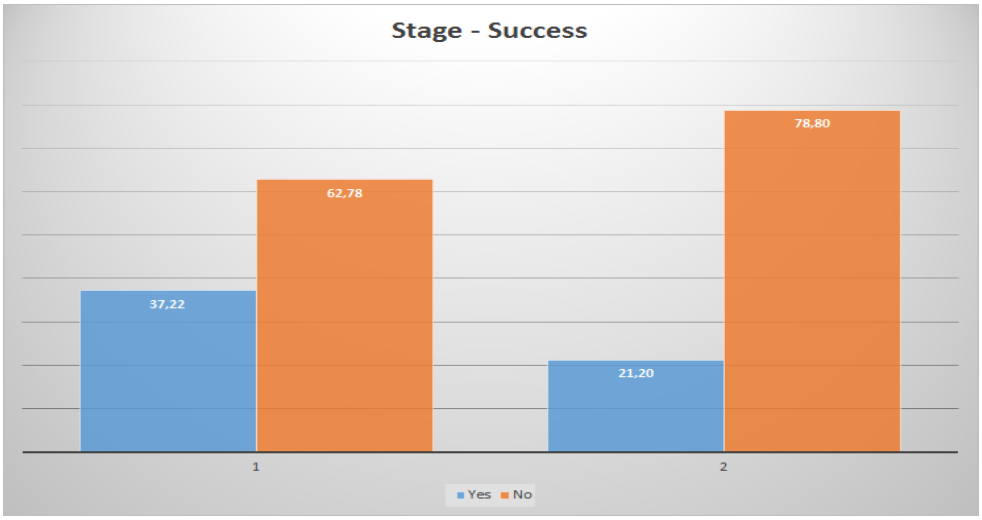

Figure 4.25: Ratio of Success on Second Semester by Enrollment Stage

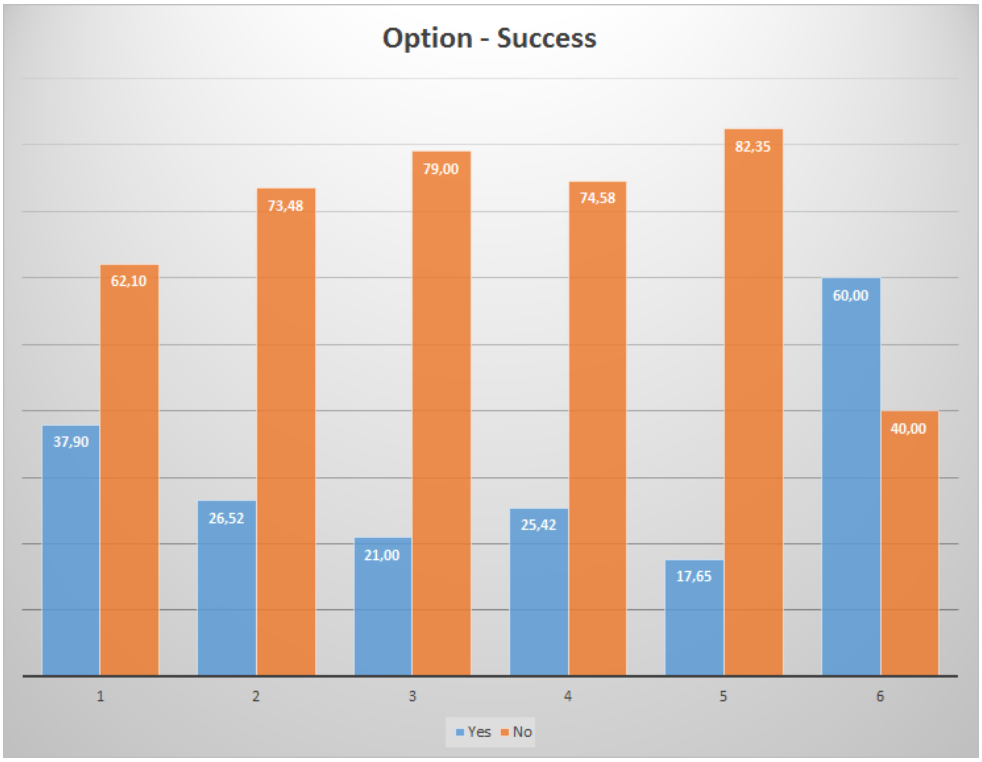

Figure 4.26: Ratio of Success on Second Semester by Enrollment Option

On figure 4.27 we can see that, similarly to what happened on the first semester, the enrollment year does not appear to be highly related to academic success, which again reinforces the need for feature selection, while also showing that the generated model should work for different years.

Figure 4.29 shows the ratios of success for the various average grades of the first semester (the distribution of these grades can be seen in figure 4.28). From this figure we can see that there is a very strong correlation, and that this attribute is likely to be one of the most impactful ones.

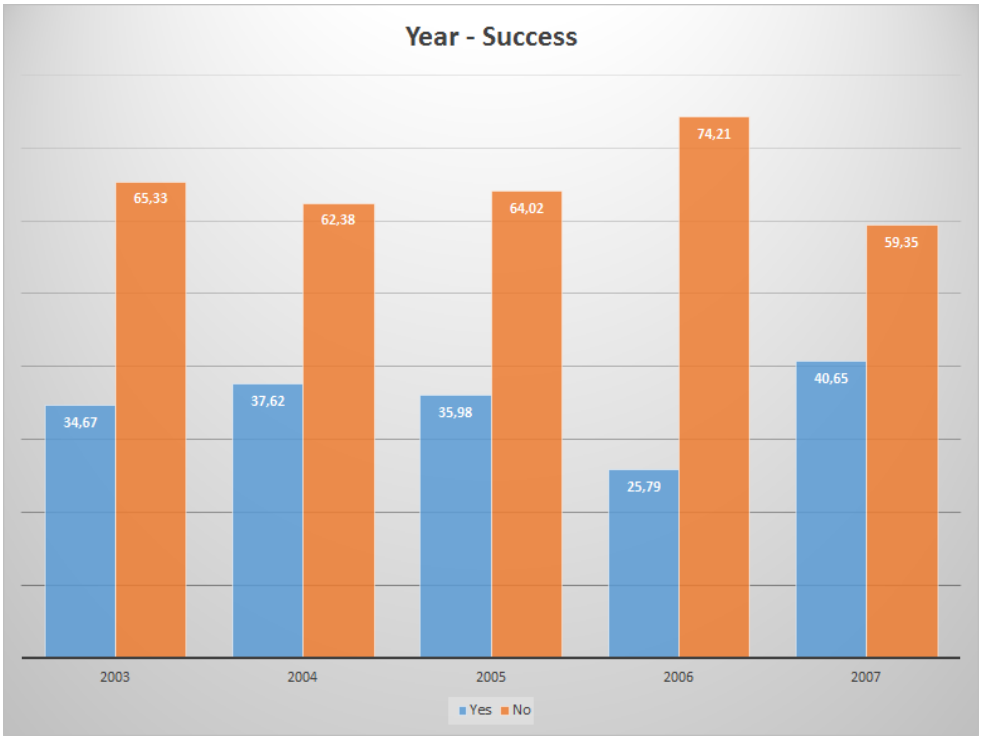

Figure 4.27: Ratio of Success on Second Semester by Enrollment Year

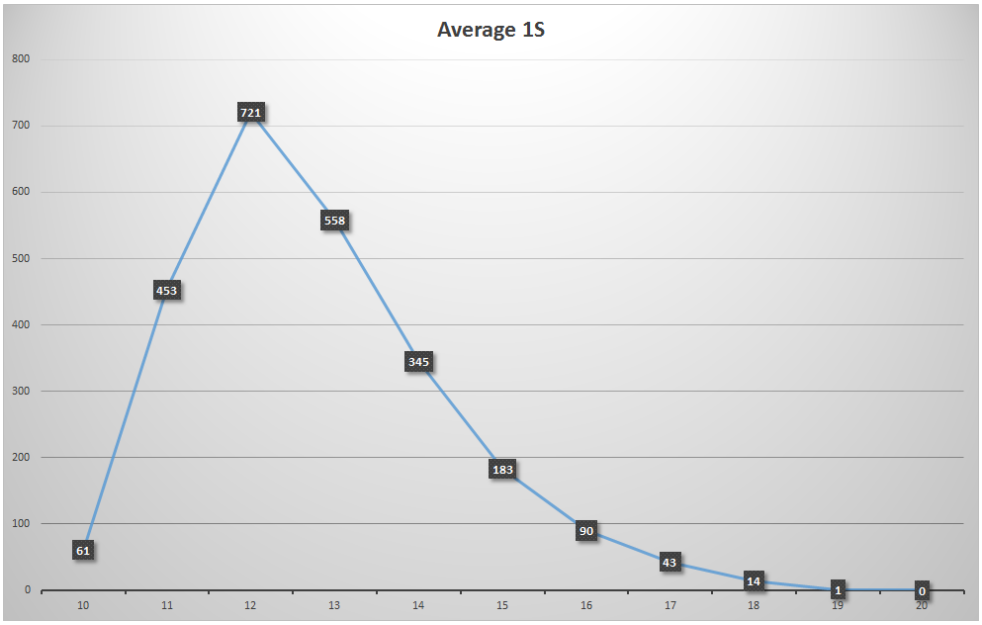

Figure 4.28: Distribution of Average Grades on First Semester

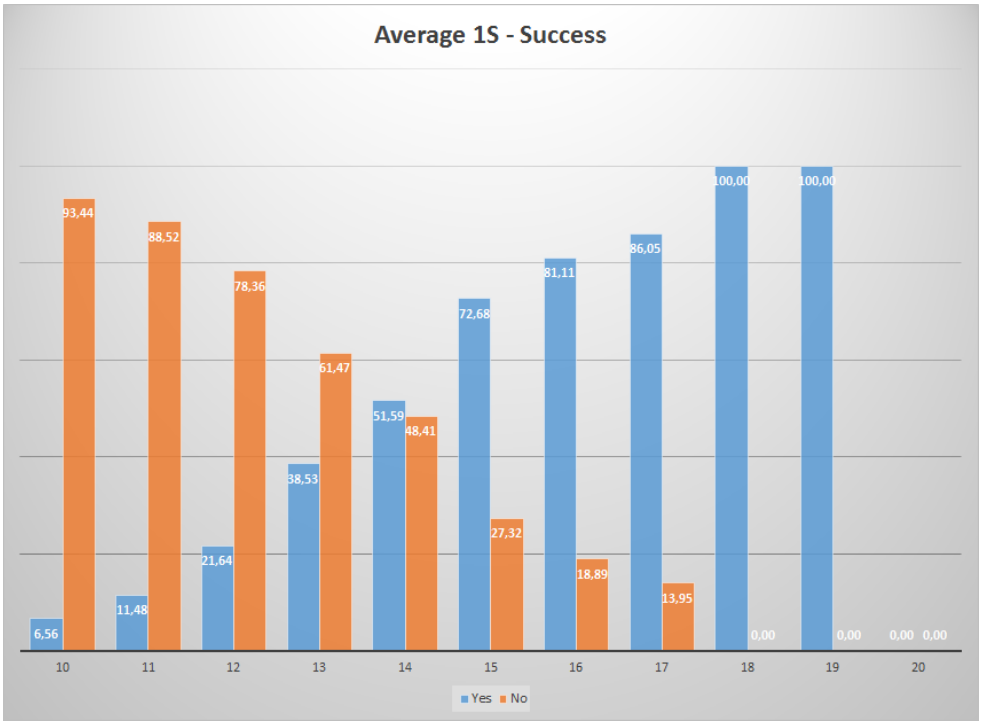

Figure 4.29: Ratio of Success on Second Semester by First Semester Average Grade

On figure 4.30 we can clearly see that academic success on the second semester (defined as completing at least 25 ECTS) is clearly related to the number of ECTS completed on the first semester, with the likelihood of achieving success increasing the more credits one completed on the first semester. The presence of a couple of outliers that achieved success despite having completed almost none in the first suggests that performing outlier detection and elimination might improve the overall performance of the model.

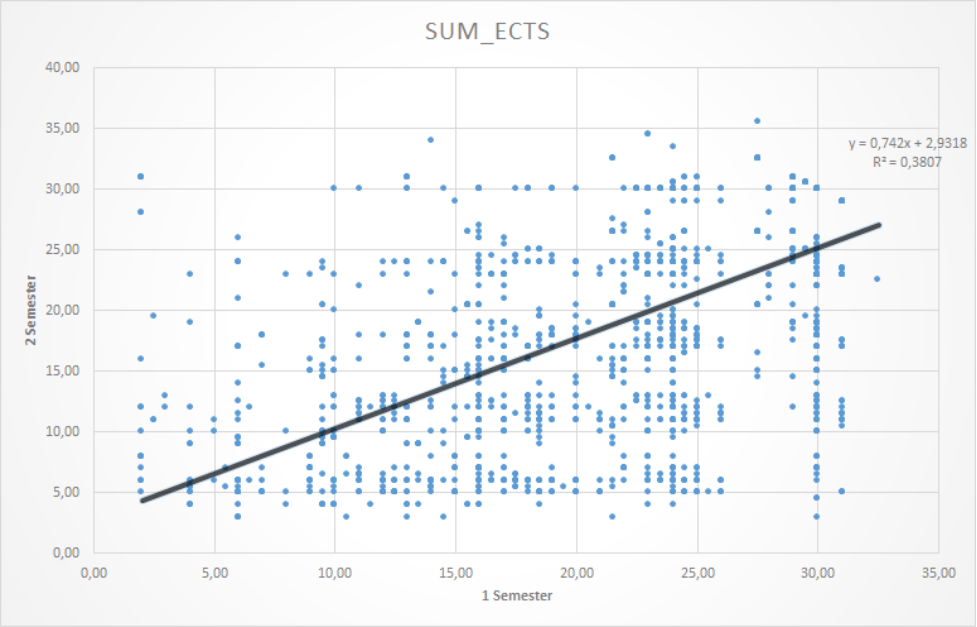

Figure 4.30: Sum of ECTS of First and Second Semester

We can therefore conclude that the information regarding the performance on the first semester, coupled with either the enrollment average grade or the high school average grade, should be the most impactful attributes for this model. We also identified the need for the application of resampling, feature selection and outlier detection techniques.

4.4 Overall success

As mentioned in the previous chapter, for measuring the overall success, we calculated a performance value by dividing the product of average grade of the student obtained at the end of his/her academic path and the number of ECTS s/he obtained approval in by the total amount of ECTS s/he enrolled in. A clustering algorithm (K-Means) was then used in order to identify how these values naturally grouped themselves, with the characteristics of these clusters leading to the performance levels seen in table 4.1. It is important to note that our goal here is not to determine the exact performance value of a student, but rather to be able to identify into which performance group s/he fits into, as determined by the grouping of the performances of his/her peers.

Performance Level	Range
A	>14.5
B	[11.5 - 14.5[
C	[9 - 11.5[
D	[6 - 9[
E	<6

Table 4.1: Performance Levels for Overall Success

We can see in figure 4.31 that this resulted in a set of balanced classes, as the levels follow the natural groupings of the data.

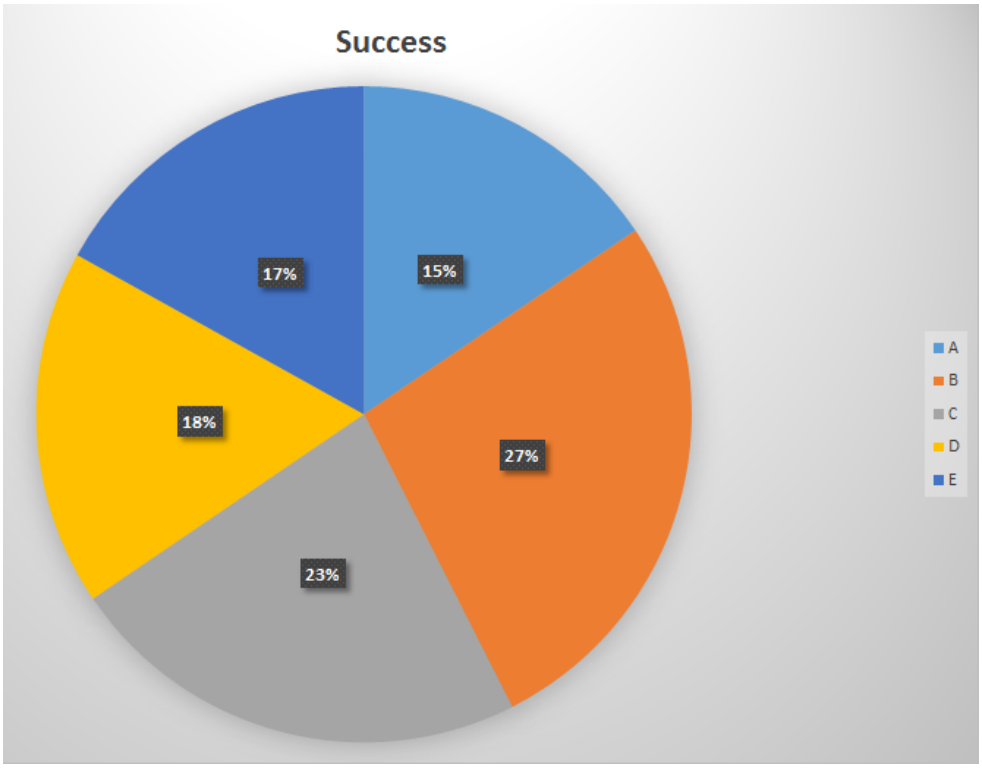

Figure 4.31: Ratio of Overall Success Levels

In figure 4.32 we can see that sex continues not to have an impact on the academic performance of the students, and continues to point to feature selection being important.

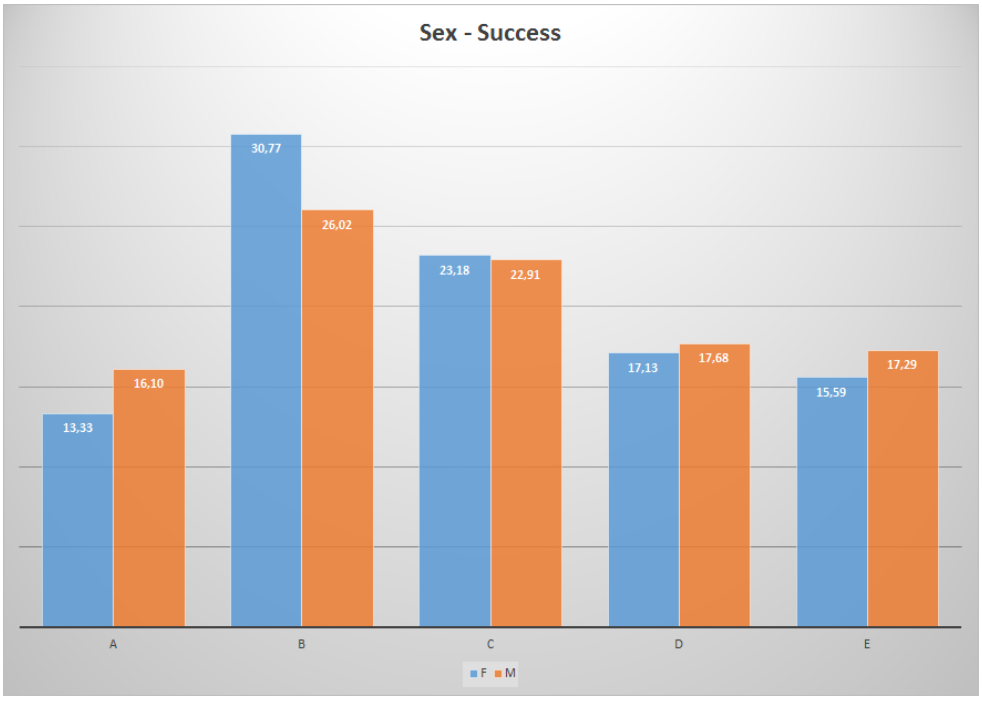

Figure 4.32: Ratio of Overall Success Levels by Sex

Figure 4.33 shows that the type of school continues to have no influence on the overall academic performance. This is to be expected, as it had no impact on either the first or second semester. Similarly, the degree the student enrolled in generally does not appear to have much impact, despite some exceptions like LEMG (figure 4.34).

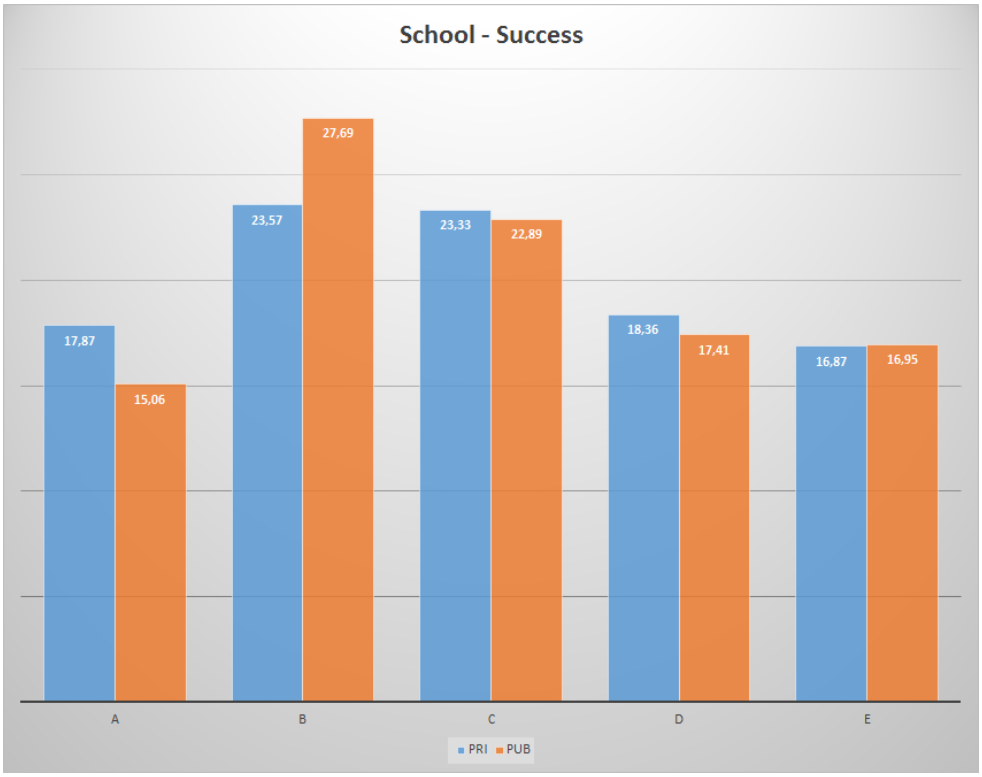

Figure 4.33: Ratio of Overall Success Levels by School Type

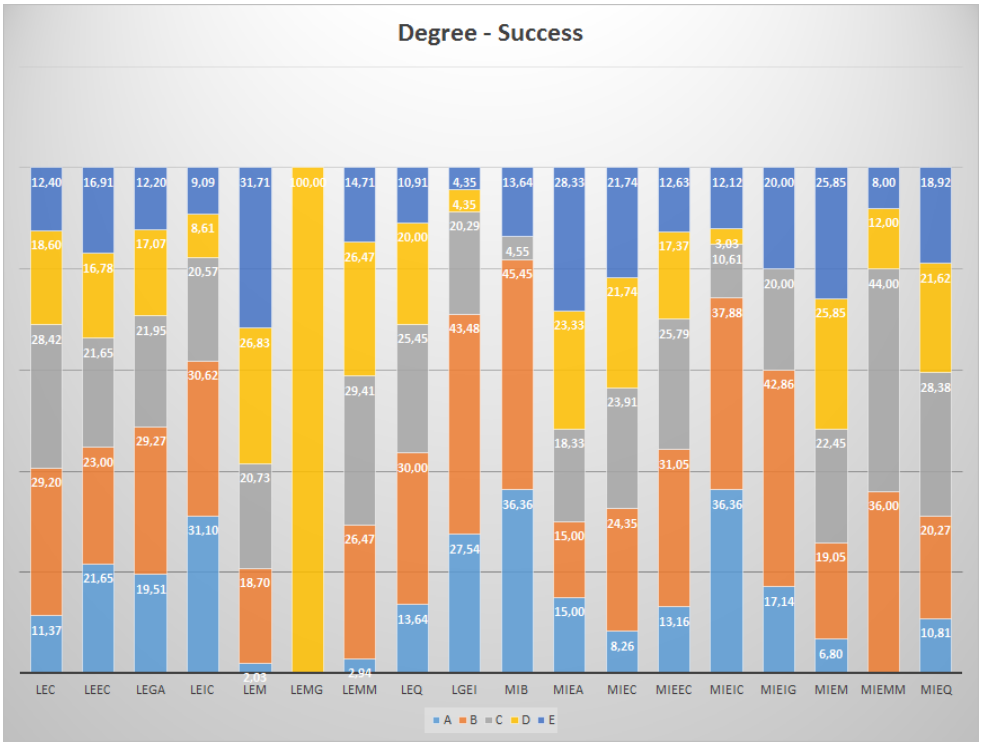

Figure 4.34: Ratio of Overall Success Levels by Degree

The change from a binary prediction class (successful/unsuccessful) into a polynomial class (five different levels of success) contributed to limiting predictive impact of the enrollment average grade and the high school average grade to the instances of very high or very low grades, as can be seen in figures 4.35 and 4.36.

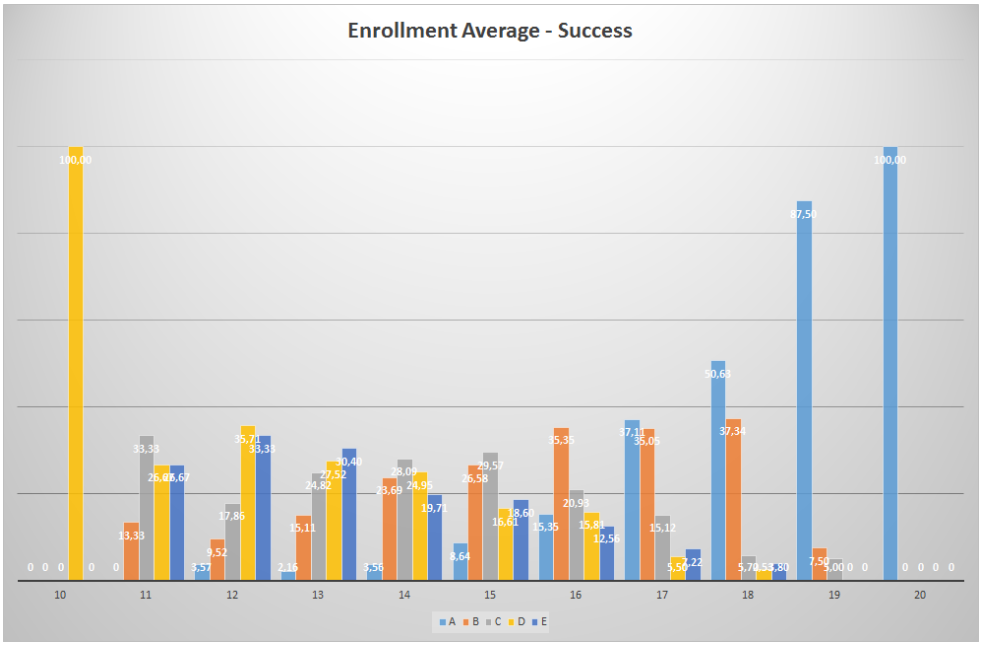

Figure 4.35: Ratio of Overall Success Levels by Enrollment Average

In figure 4.37 we can see that students who enrolled in the second stage don't usually have the highest level of performance, but for all other levels, there doesn't seem to be any difference. As for the enrollment option, while there seems to be some correlation, with the fifth option surprisingly tending towards the second-highest level of performance, it does not otherwise appear to be enough to have a high predictive weight, especially considering how predominantly the first two options feature among the dataset.

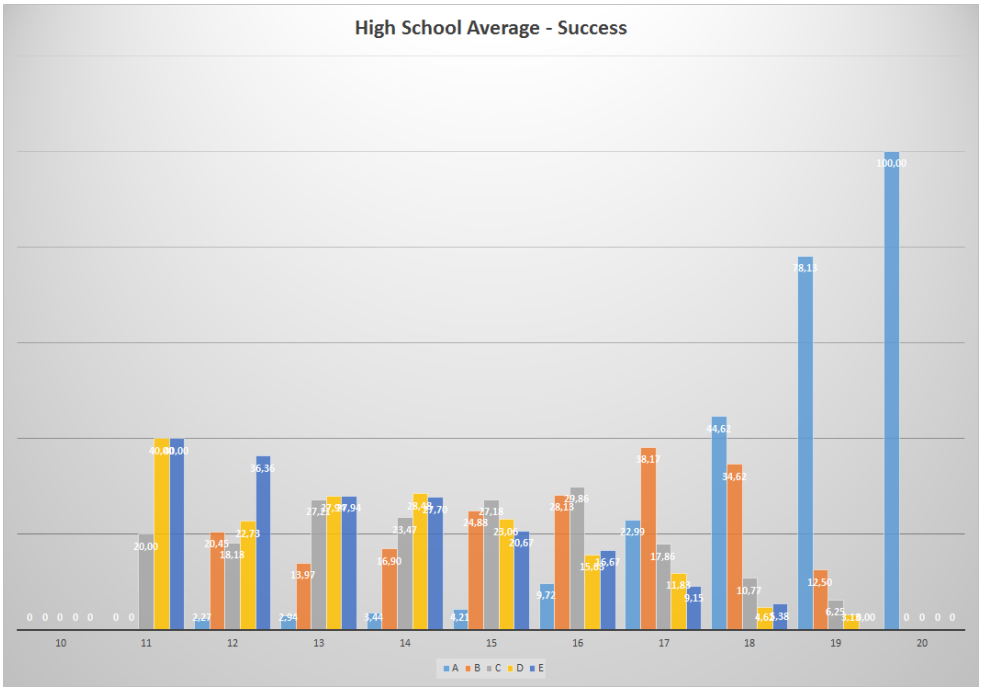

Figure 4.36: Ratio of Overall Success Levels by High School Average

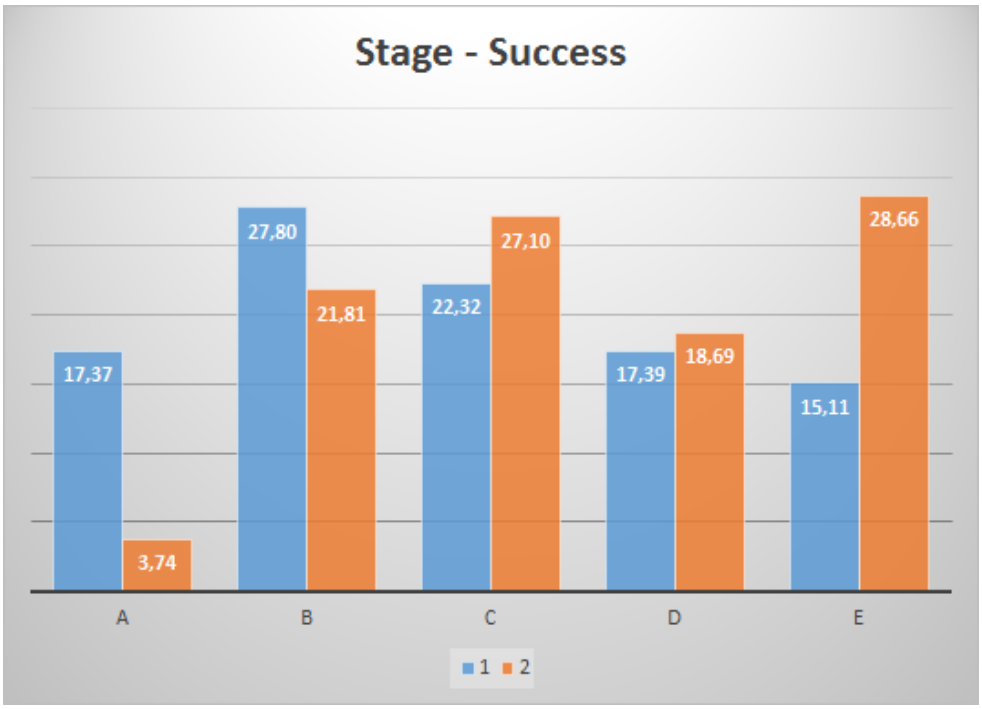

Figure 4.37: Ratio of Overall Success Levels by Enrollment Stage

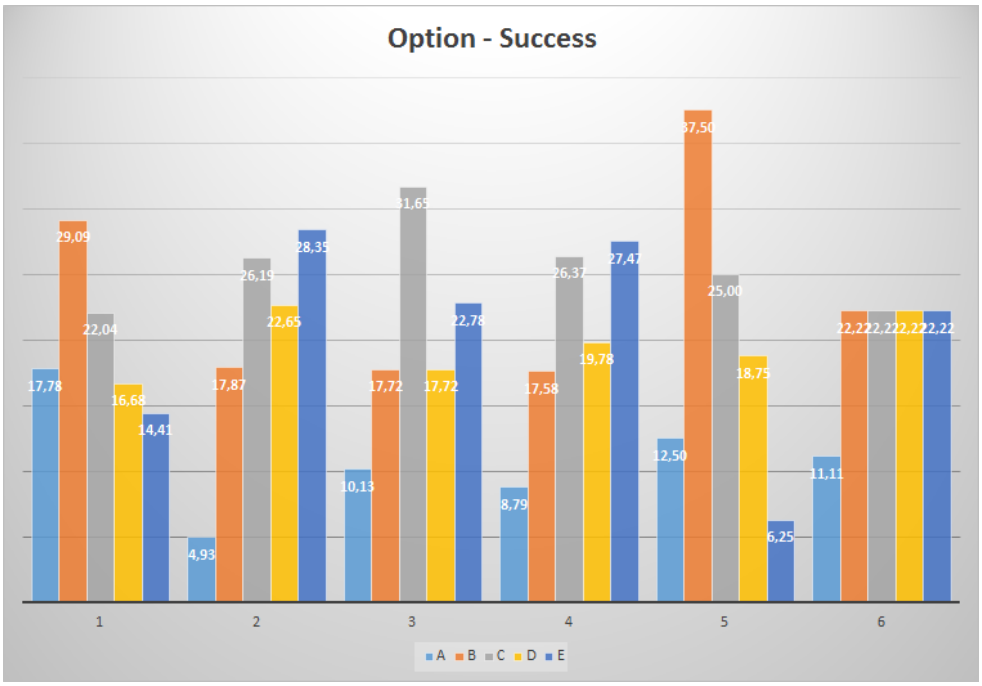

Figure 4.38: Ratio of Overall Success Levels by Enrollment Option

Once again, the enrollment year (figure 4.39) does not appear to have any relevant impact on academic performance, and continues to suggest both the use of feature selection to reduce noise and that it is possible to generalize these predictive models for other years.

In figure 4.40 we can observe that the average grade of the first semester seems to have the highest correlation of all the attribute analyzed so far for this model, with very clear biases for almost all of its possible values.

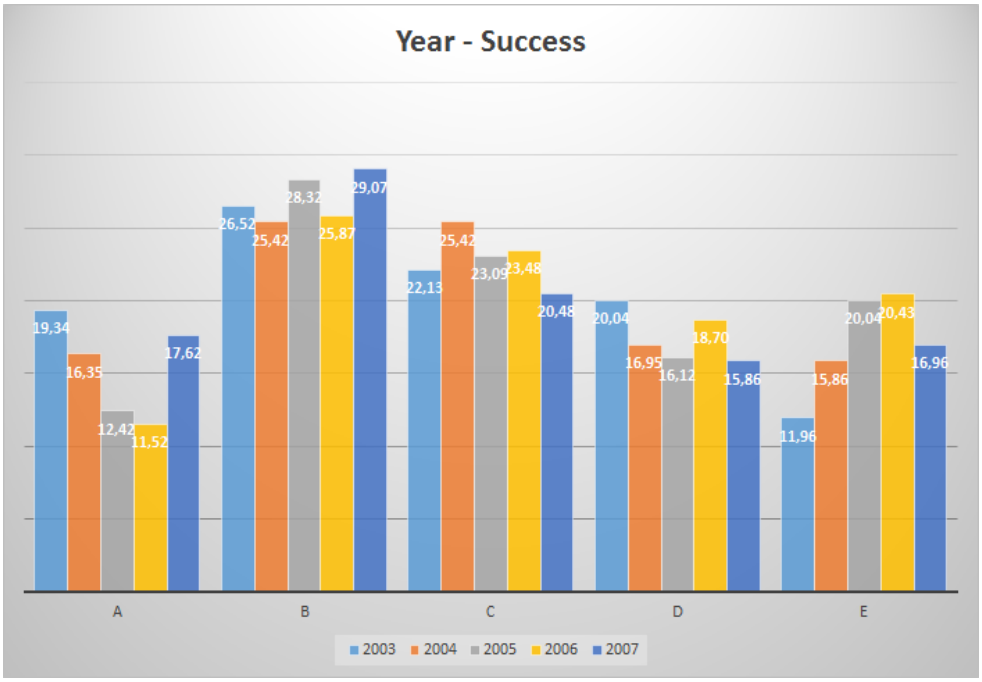

Figure 4.39: Ratio of Overall Success Levels by Enrollment Year

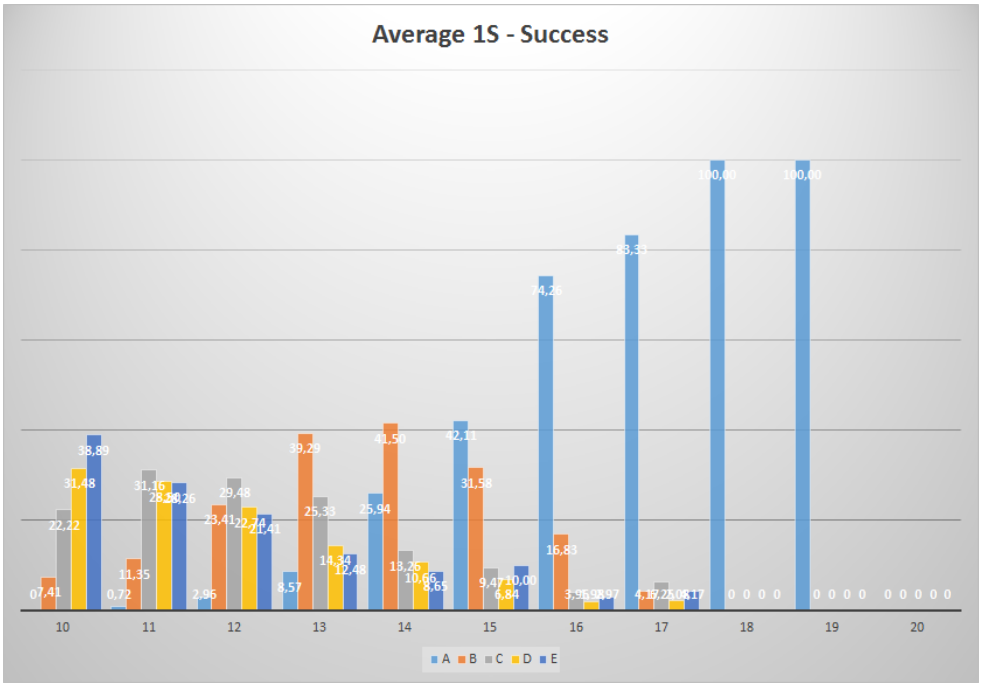

Figure 4.40: Ratio of Overall Success Levels by First Semester Average Grade

The average grades of the second semester appear to lead to have the same type of impact, and approximately the same tendencies, as the average of the first semester, so it is very likely that the models will use only of them, due to their high correlation (see figure 4.42). The distribution of the average grades of the second semester can be seen in figure 4.41, which presents us with a mean grade of 12.85, with a standard deviation of 5.93.

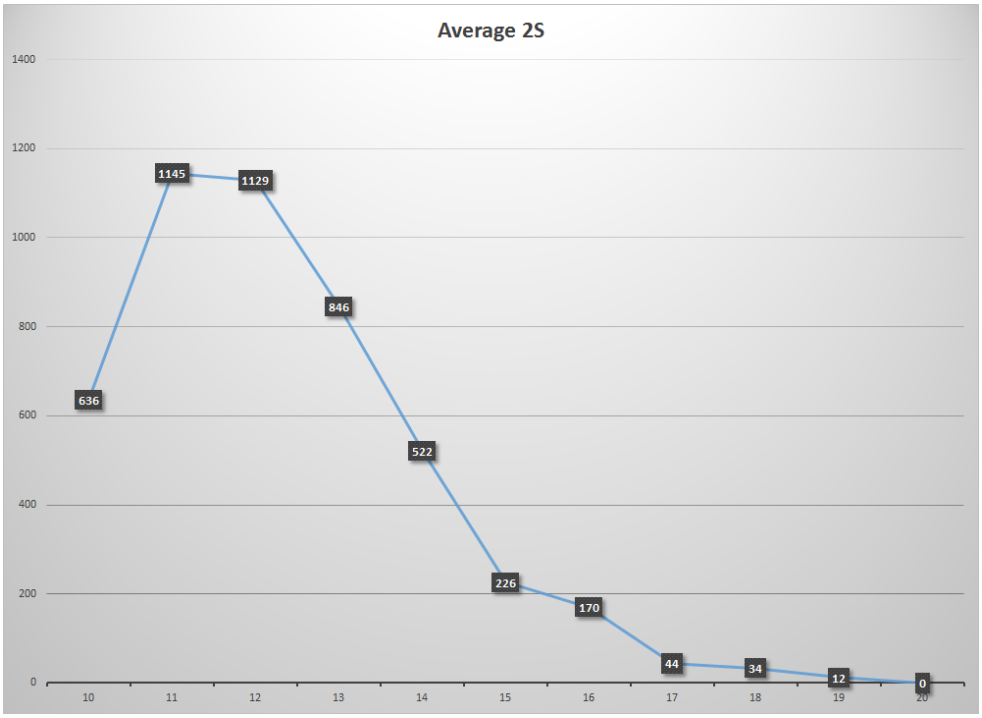

Figure 4.41: Distribution of Average Grades on Second Semester

Lastly, there seems to be a very high correlation between the sum of credits of both the first and the second semesters and the performance value, with the correlation being stronger for the second semester (this is shown in figures 4.43 and 4.44). This seems to indicate that these values will have a very high predictive impact on the model. Also, similarly to what happened on the previous model, some outliers are present, with the first semester showing a higher degree of them.

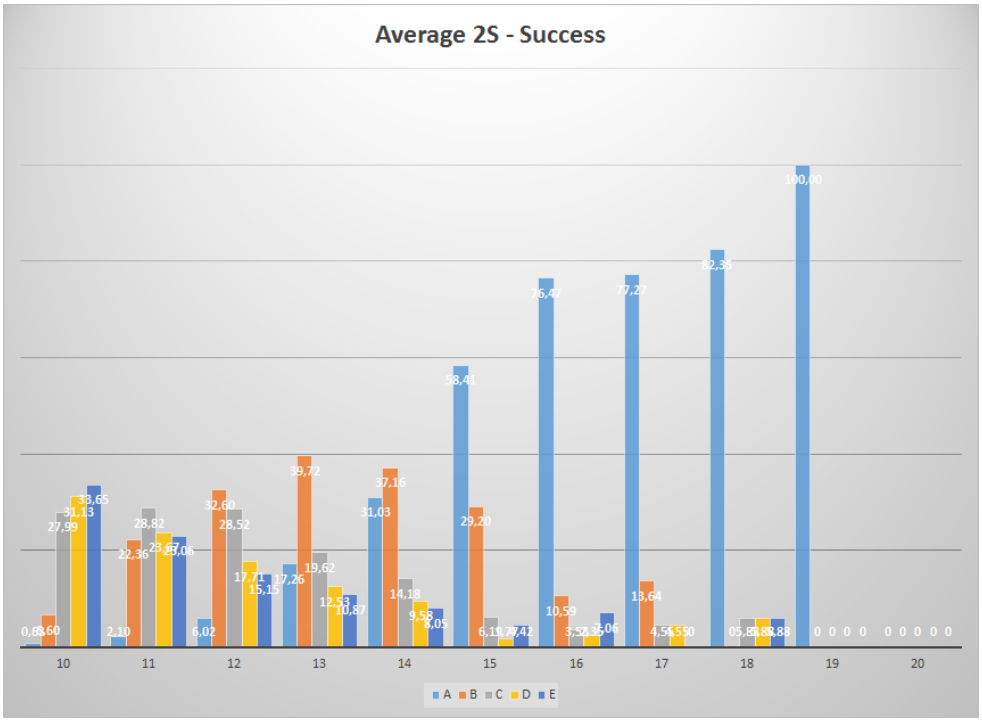

Figure 4.42: Ratio of Overall Success Levels by Second Semester Average Grade

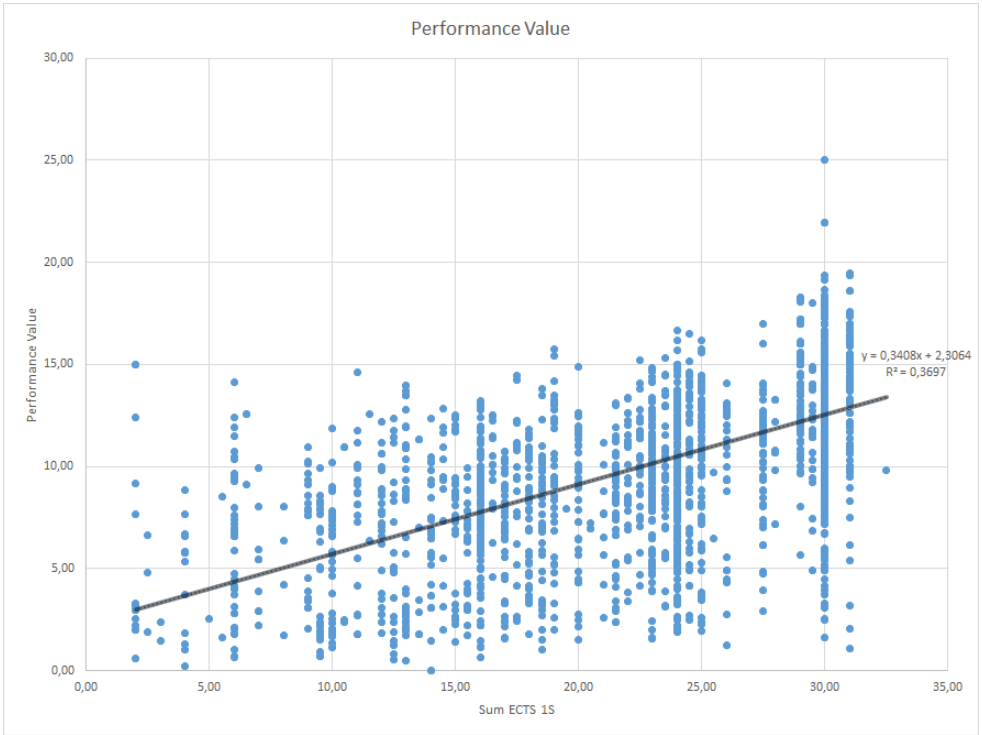

Figure 4.43: Performance Values and Sum of ECTS of First Semester

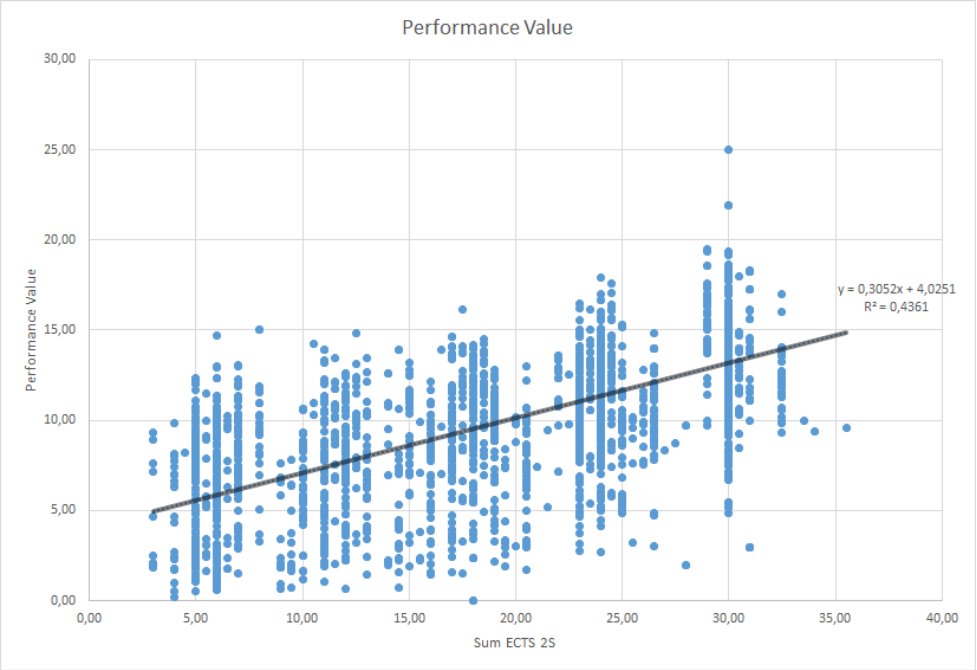

Figure 4.44: Performance Values and Sum of ECTS of Second Semester

From this analysis, we can conclude that the sum of credits from the second semester and one of the average grades of a semester (possibly the first, to reduce bias from using the second semester's sum of credits) are the attributes most likely to have a preponderant role in the predictive model. Furthermore, we were able to identify the need for feature selection and outlier detection to improve the data quality for this model.

4.5 Data Preparation

A big part of achieving good results in Data Mining processes is having good quality data. The presence of incorrect or corrupted values, or even of other types of noise, in the dataset, can create biases that lead to misleading results. Therefore, data preparation is of crucial importance in any Data Mining project. Data preparation consists of a series of tasks that perform modifications to a dataset in order to improve its quality.

One of the most basic data preparation tasks is the removal or modification of incorrect or missing values. This was performed on entries about the students' parents and about their addresses, as mentioned previously.

Datasets sometimes contain outliers, instances which do not conform to the patterns formed by the other elements in the dataset. These instances can introduce noise into the model, but care should be taken not to incur relevant loss of information from their elimination. However, within the scope of this project, since the aim is to analyze the typical student (ignoring, for example, students who enrolled in a non-regular contingent), the choice was made to eliminate outliers, calculating them based on euclidean distance. The number of outliers eliminated was, however, kept below 1% of the total instances in all cases, to minimize any loss of information that might come from it.

Another data preparation task that we used in this project was resampling. The analysis of success on the first year showed that it had an unbalanced division. This could lead to misleading results, as the algorithm could achieve decent results just by favoring the majority class, while not truly being able to identify its elements. By oversampling the minority class, that is by recreating the dataset while repeating some of the instances of that class, we were able to reach a much more balanced final dataset that provides a better generalization capability. This was helped by the fact that even with oversampling, the number of instances was small enough that there was no need to opt for a hybrid approach of both oversampling and undersampling in order to maintain a decent performance. This allowed us to avoid any loss of information that inevitably comes from undersampling (selecting less than the total instances available from one class).

Lastly, a dataset often contains attributes that are either redundant or irrelevant, and that can therefore be removed without much loss of information. This can not only enhance generalization but also lead to simpler models, making them easier to interpret. This process of selecting relevant attributes for use in a model is called feature selection. During our exploratory analysis, we were able to identify various attributes that hinted at the need for feature selection, such as the year of enrollment of the student. We tackled this issue by determining the information gain of each attribute, and then removing the attributes that were under a certain threshold (0.05 for the first model, 0.08 for the others).

Chapter 5

Results

In order to understand the obtained results, it is important to understand the parameters used for them. In this regard, Random Forest used 100 trees with a K equal to approximately the square root of the total number of attributes, while C4.5 was developed as an unprunned tree. For SVM, a complexity constant of 1 was used, with a tolerance parameter of 0.001. Logistic models are also fitted to SVM outputs, which allows for the calculation of relevant AUC values, as Weka's implementation of SMO changes the output values to the extremes without this parameter (meaning positive classifications take the value of 100 and negative ones the value of 0). These parameters were used in all three developed models.

The first model developed uses information available at the start of the student's first year to predict success at the end of his/her first semester. Following feature selection, the attributes that were used by the model were:

- mother's education level
- father's education level
- mother's job
- father's job
- enrollment option
- enrollment stage
- degree that the student enrolled in
- high school average grade
- enrollment exams average grade
- enrollment average grade

- enrollment exams

The exception to the above list lies with the Random Forest algorithm, which did not go through feature selection, leading to it including these attributes as well:

- school type
- marital status
- enrollment year
- sex

As mentioned previously, the data for these models also underwent outlier detection and removal, with 2459 instances remaining afterward (a total of 10 outliers removed). It's important to note, however, that this dataset was then subjected to oversampling, resulting in a total of 3145 instances.

The performance of the four algorithms for this model can be seen in table 5.1.

Algorithm	Accuracy	Specificity	AUC
RandomForest (I = 100, K = 4)	80.83%	93.94%	94.00%
j48 (unprunned)	83.85%	85.41%	88.40%
Bayes	77.87%	79.51%	84.70%
SVM	83.31%	87.76%	89.90%

Table 5.1: Performance of first model

From the above table we can see that Random Forest and SVM produce the best AUC, with Random Forest having a worse Accuracy but having a better coverage of the unsuccessful students, as is indicated by the higher Specificity. Meanwhile, j48 has the highest Accuracy and provides us with information regarding the attributes which have a higher weight in the academic success of FEUP's students, as seen in figure 5.1. Naive Bayes produces the worst results of the four.

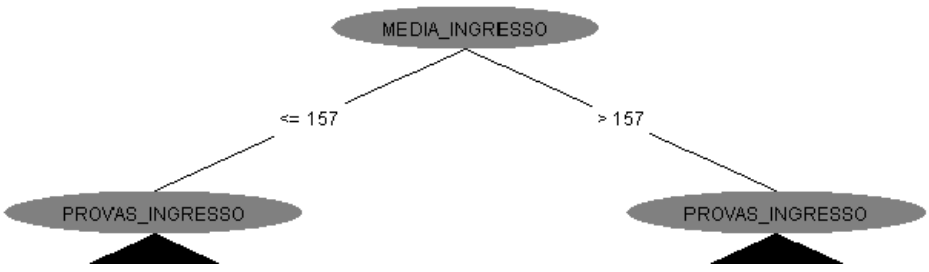

Figure 5.1: Top Levels of the Decision Tree of First Model

By analyzing figure 5.1 we conclude that the variables MEDIA_INGRESSO (the enrollment average grade) and PROVAS_INGRESSO (the exams performed for enrollment) are the attributes with the most weight, as they are the ones where the first divisions of the tree occur. This is in

accordance with the exploratory analysis performed on the previous chapter, where we saw that either the enrollment average or the high school grade were very likely to feature as key attributes.

The second model developed uses information available at the end of the first semester of the student's first year to predict success at the end of his/her second semester. Following feature selection, the attributes that remained in use by this model were:

- mother's education level
- father's education level
- mother's job
- father's job
- enrollment stage
- degree that the student enrolled in
- high school average grade
- enrollment exams average grade
- enrollment average grade
- enrollment exams
- number of college exams for approval
- number of college exams for grade improvement
- average grade on the first semester
- number of ECTS completed on the first semester

Once again, the exception to the above list lies with the Random Forest algorithm, which did not go through feature selection, and therefore included these attributes as well:

- school type
- marital status
- enrollment year
- enrollment option
- sex

Algorithm	Accuracy	Specificity	AUC
RandomForest (I = 100, K = 5)	90.21%	94.59%	97.60%
j48 (unprunned)	85.68%	94.39%	95.30%
Bayes	82.21%	83.50%	89.90%
SVM	89.23%	91.18%	95.80%

Table 5.2: Performance of second model

This data also underwent outlier detection and removal, with a removal of 15 outliers, leaving us with 2454 instances. After the resulting dataset underwent oversampling, we were left with a total of 3136 instances.

The performance of the four algorithms for this model can be seen in table 5.2.

In this model, Random Forest again produces the top results, while Naive Bayes is still the worst performing algorithm. j48 continues to provide us with information regarding the attributes which have a higher weight in the academic success of FEUP's students, as seen in figure 5.2.

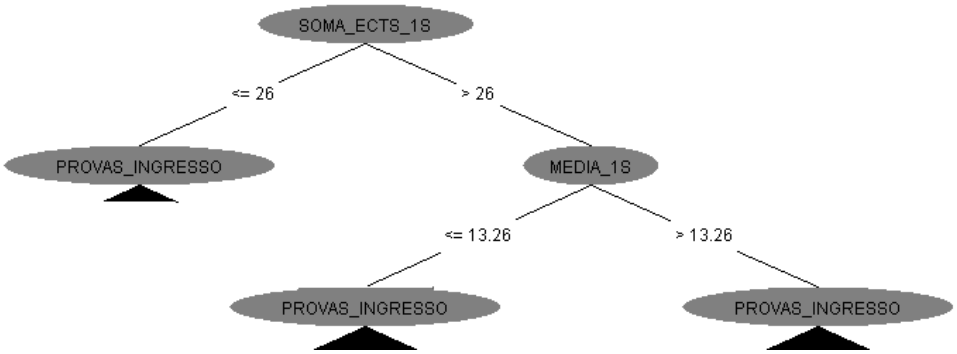

Figure 5.2: Top Levels of the Decision Tree of Second Model

By analyzing figure 5.2 we can see that SOMA_ECTS_1S (the amount of credits the student concluded in his/her first semester), PROVAS_INGRESSO (the exams performed for enrollment) and MEDIA_1S (the average grade of the student on his/her first semester) are the attributes with the most weight for this second model, and also that the amount of credits concluded is more relevant for identifying successful/unsuccessful students than their average grade. This is also in accordance with the exploratory analysis performed on the previous chapter, where we concluded that these three attributes were the most likely ones to have a key role in the prediction model, though it is interesting that for low credit sums, the average does not feature prominently in the decision tree.

Lastly, the third model developed uses information available at the end of student's first year to predict success at the end of his/her degree or at 2015, whichever came first. This model also underwent feature selection, with the resulting attributes being:

- degree that the student enrolled in
- high school average grade

- enrollment average grade
- enrollment exams
- number of college exams for approval
- average grade on the first semester
- number of ECTS completed on the first semester
- average grade on the second semester
- number of ECTS completed on the second semester

The Random Forest algorithm continued to be the exception to the above list, as it did not go through feature selection, and so included these attributes in addition to the above:

- mother's education level
- father's education level
- mother's job
- father's job
- school type
- marital status
- enrollment exams average grade
- enrollment year
- enrollment stage
- enrollment option
- sex
- number of college exams for grade improvement

Outlier detection and removal were also performed on this model, with a total of 2459 instances remained afterward (a total of 10 outliers removed). Contrary to what happened in the previous two models, since no resampling occurred in this model, this was performed over the initial dataset.

Once again, the performance of the four algorithms for this model can be seen in table 5.3.

Random Forest continues to produce some of the best results out of the four algorithms, with a very high performance. j48 once again provides us with information regarding the attributes which have a higher weight in the academic success of FEUP's students, as seen in figure 5.3.

Algorithm	Accuracy	Recall	Precision
RandomForest (I = 100, K = 5)	96.54%	95.87%	96.90%
j48 (unprunned)	91.79%	91.54%	91.82%
Bayes	75.11%	74.14%	74.44%
SVM	92.15%	91.22%	91.97%

Table 5.3: Performance of third model

By once again analyzing the decision tree (presented in figure 5.3) we can see that SOMA_ECTS_2S (the amount of credits the student concluded in his/her second semester), MEDIA_2S (the average grade of the student on his/her second semester) and PROVAS_INGRESSO (the exams performed for enrollment) are the attributes with the most weight. These attributes were identified in the previous chapter as having a strong correlation with overall academic success, though the exploratory analysis performed led us to believe that the average used would be the one from the first semester.

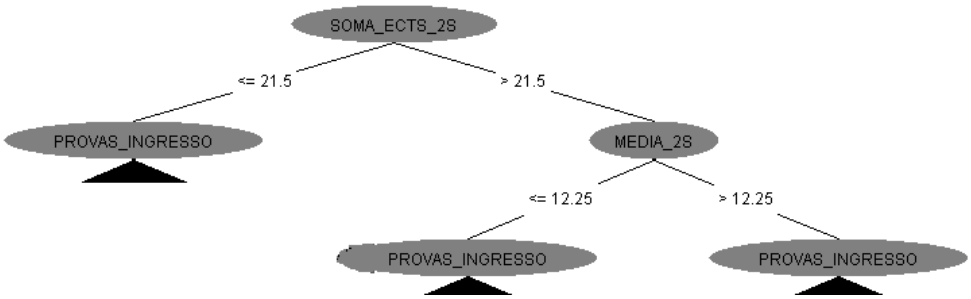

Figure 5.3: Top Levels of the Decision Tree of Third Model

Chapter 6

Conclusions and Further Work

6.1 Conclusions

In order to reduce time to degree completion colleges need to be able to identify successful and unsuccessful students in the beginning of their academic path, as well as to identify what appears to have a greater impact on that distinction. For this purpose, three prediction models were developed: one that predicts success on the first semester (defined as completing at least 25 ECTS on that semester) with the data available at the time of enrollment, one that predicts success on the second semester (also defined as completing at least 25 ECTS on that semester) with the data available at the end of the first semester, and a third that would predict overall academic success (defined by five performance levels related to the ratio between the student's average grade and the number of ECTS he enrolled in) with the information available at the end of the first year.

In regards to all three models, the balance of the work presented here is extremely positive, as not only do the indicators for the accuracy of our models commonly reach upwards of 80% (and in some cases, even 90%), but we were also able, through exploratory analysis of the data and through the analysis of the decision trees generated, to identify the most impactful attributes on each of the three models. This analysis showed that the enrollment average and the entrance exams were the attributes with the most correlation to success at the first semester. In the other two models, however, the information about the first and then the second semesters (number of ECTS completed and average grade) takes the place of the enrollment average. Nonetheless, it is important to note that while the information about the second semester would replace the one about the first one, the entrance exams continue to have a big impact in all three models). Due to a lack of information, however, this analysis did not extend itself to which specific courses had the most predictive impact.

It's also interesting to note how the Random Forest algorithms is the one which consistently results in the best performance, even though the variation in performance is not too large, except in the third model (the only multinomial classification one), where Bayes Network shows a significantly worse accuracy than the other three algorithms.

6.2 Future Work

The future work for the project developed relies on improving the models and integrating them into a management platform. One of the caveats of the models developed for this dissertation is that they only take into account full-time students who enrolled in the regular contingent. Having proved the prediction ability of these models, the next step would be to expand these models to take into account all types of students, so as to better serve a more diverse population. Another limitation of these models resulted from the lack of information regarding the addresses of students and whether they were beneficiaries of grants, as well as lack of information regarding academic involvement and specific courses. The first variables, while present in the dataset, are unreliable or have a lot of missing entries, while the latter two are not present in it at all. On the long term, changing how FEUP collects this information, in order to enable its integration in these models, could be a great boon to the quality of the predictions that result from them. Lastly, while the work produced herein gives college administration useful information about student behavior regarding academic performance, an integrated platform that would allow to make predictions and identify certain segments of the student population could be even more useful. Though such a platform was outside the scope of this dissertation, it is the opinion of the author that, following an inclusion of information on specific courses on these models, this would be the most gainful way of continuing the work started in this dissertation, because of the direct use for college administrators.

References

Agrawal, R., Imieliński, T., & Swami, A. (1993, June). Mining association rules between sets of items in large databases. *SIGMOD Rec.*, *22*(2), 207–216. Retrieved from http://doi.acm.org/10.1145/170036.170072 doi: 10.1145/170036.170072

Almeida, L. S., & Cruz, J. F. A. (2010). Transição e adaptação académica: reflexões em torno dos alunos do 1º ano da universidade do minho.

Asif, R., Merceron, A., & Pathan, M. K. (2014). Predicting student academic performance at degree level: a case study. *International Journal of Intelligent Systems and Applications (IJISA)*, *7*(1), 49.

Asif, R., Merceron, A., & Pathan, M. K. (2015). Predicting student performance at degree level: a case study. *International Journal of Intelligent Systems and Applications*. doi: 10.5815/ijisa.2015.01.05

Baker, R., et al. (2010). Data mining for education. *International encyclopedia of education*, *7*, 112–118.

Baker, R. S., Corbett, A. T., & Wagner, A. Z. (2006). Human classification of low-fidelity replays of student actions. In *Proceedings of the educational data mining workshop at the 8th international conference on intelligent tutoring systems* (pp. 29–36).

Baker, R. S., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. *JEDM-Journal of Educational Data Mining*, *1*(1), 3–17.

Beck, J. E., & Mostow, J. (2008). How who should practice: Using learning decomposition to evaluate the efficacy of different types of practice for different types of students. In *Intelligent tutoring systems* (pp. 353–362).

Chau, V. T. N., & Phung, N. H. (2013, Nov). Imbalanced educational data classification: An effective approach with resampling and random forest. In *Computing and communication technologies, research, innovation, and vision for the future (rivf), 2013 ieee rivf international conference on* (p. 135-140). doi: 10.1109/RIVF.2013.6719882

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning*, *20*(3), 273–297.

García, E., Romero, C., Ventura, S., & de Castro, C. (2011). A collaborative educational association rule mining tool. *The Internet and Higher Education*, *14*(2), 77–88.

Gong, Y., Rai, D., Beck, J. E., & Heffernan, N. T. (2009). Does self-discipline impact students' knowledge and learning?. *International Working Group on Educational Data Mining*.

Hajizadeh, N., & Ahmadzadeh, M. (2014). Analysis of factors that affect students' academic performance-data mining approach. *arXiv preprint arXiv:1409.2222*.

Hodge, V., & Austin, J. (2004). A survey of outlier detection methodologies. *Artificial Intelligence Review*, *22*(2), 85-126. Retrieved from http://dx.doi.org/10.1007/s10462-004-4304-y doi: 10.1007/s10462-004-4304-y

Huebner, R. A. (2013). A survey of educational data mining research.

Kabakchieva, D. (2013). Predicting student performance by using data mining methods for classification. *Cybernetics and Information Technologies*, *13*(1), 61–72.

References

Kotsiantis, S. B., Zaharakis, I., & Pintelas, P. (2007). *Supervised machine learning: A review of classification techniques.*

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The annals of mathematical statistics*, 79–86.

Machado, J., Curado, A. P., et al. (2006). Percursos escolares dos estudantes da universidade de lisboa: à entrada: um retrato sociográfico dos estudantes inscritos no 1º ano.

MacKay, D. J. (2003). *Information theory, inference and learning algorithms*. Cambridge university press.

Mendes, F. (2007). O desempenho dos alunos no ensino superior politécnico perspectivado a partir da classificação no ensino secundário e na nota de candidatura. *Revista portuguesa de Pedagogia*(41-2), 29–49.

Murty, M. N., & Devi, V. S. (2011). *Pattern recognition: An algorithmic approach*. Springer Science & Business Media.

Nghe, N. T., Janecek, P., & Haddawy, P. (2007, Oct). A comparative analysis of techniques for predicting academic performance. In *Frontiers in education conference - global engineering: Knowledge without borders, opportunities without passports, 2007. fie '07. 37th annual* (p. T2G-7-T2G-12). doi: 10.1109/FIE.2007.4417993

Oskouei, R. J., & Askari, M. (2014). Predicting academic performance with applying data mining techniques (generalizing the results of two different case studies). *Computer Engineering and Applications Journal*, *3*(2), 79–88.

Pal, A. K., & Pal, S. (2013). Data mining techniques in edm for predicting the performance of students. *International Journal of Computer and Information Technology (ISSN: 2279–0764) Volume*.

Platt, J. C. (1998). *Sequential minimal optimization: A fast algorithm for training support vector machines* (Tech. Rep.). ADVANCES IN KERNEL METHODS - SUPPORT VECTOR LEARNING.

Quinlan, J. R. (2014). *C4. 5: programs for machine learning*. Elsevier.

Ranjan, J., & Malik, K. (2007). Effective educational process: a data-mining approach. *Vine*, *37*(4), 502–515.

Romero, C., & Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. *Expert Systems with Applications*, *33*(1), 135 - 146. Retrieved from http://www.sciencedirect.com/science/article/pii/S0957417406001266 doi: http://dx.doi.org/10.1016/j.eswa.2006.04.005

Steinbach, M., Karypis, G., Kumar, V., et al. (2000). A comparison of document clustering techniques. In *Kdd workshop on text mining* (Vol. 400, pp. 525–526).

Strecht, P., Moreira, J. M., & Soares, C. (2014). Educational data mining: preliminary results at university of porto.

Chapter 7

Appendix A - AUC Curves for First Model

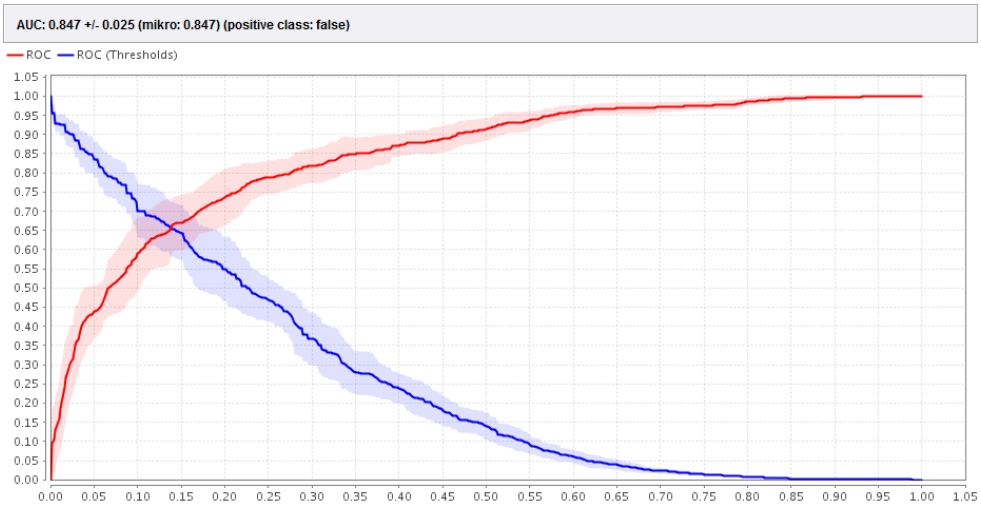

Figure 7.1: AUC Curve for Bayes (First Model)

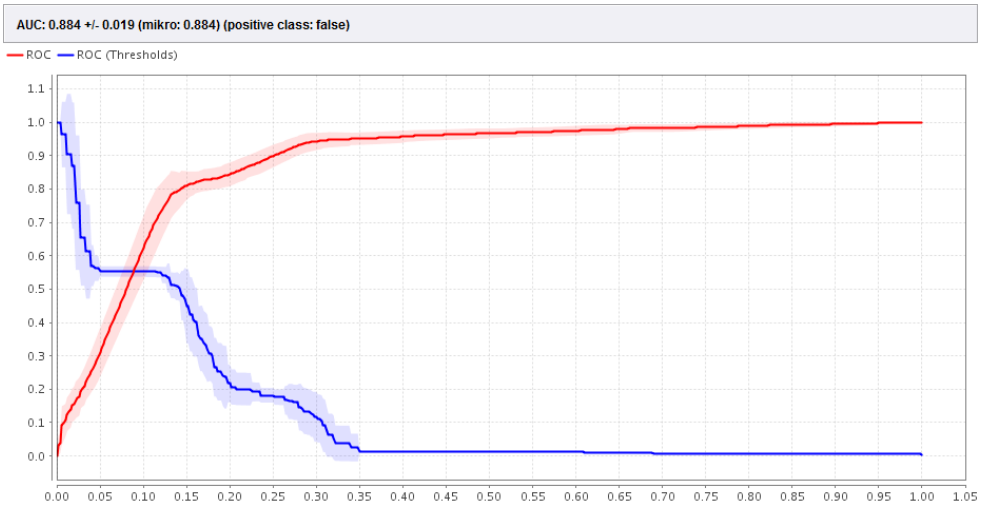

Figure 7.2: AUC Curve for j48 (First Model)

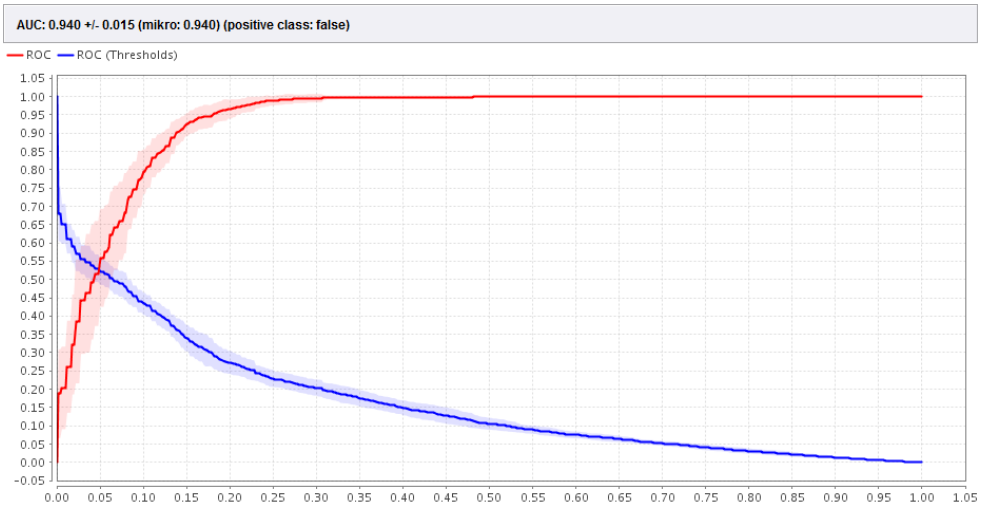

Figure 7.3: AUC Curve for Random Forest (First Model)

Appendix A - AUC Curves for First Model

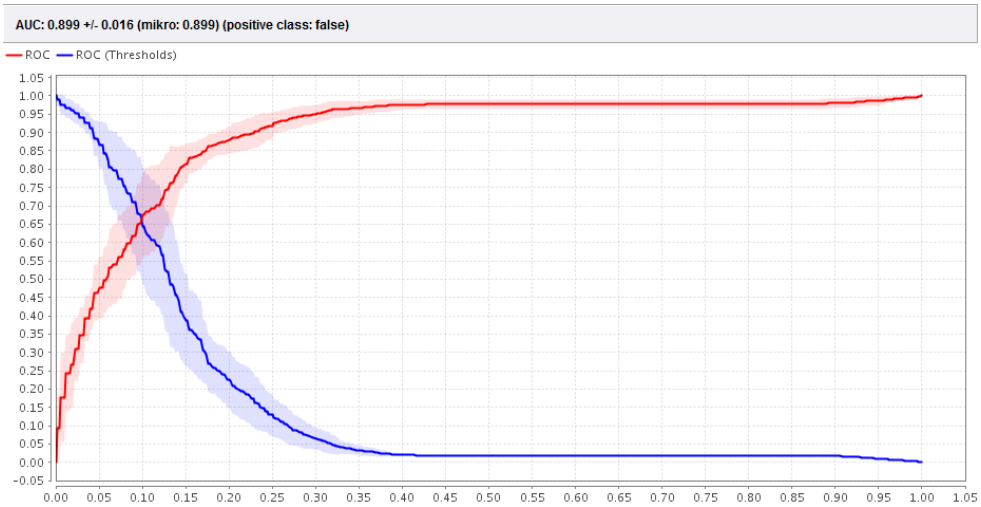

Figure 7.4: AUC Curve for SVM (First Model)

Appendix A - AUC Curves for First Model

Chapter 8

Appendix B - AUC Curves for Second Model

Figure 8.1: AUC Curve for j48 (Second Model)

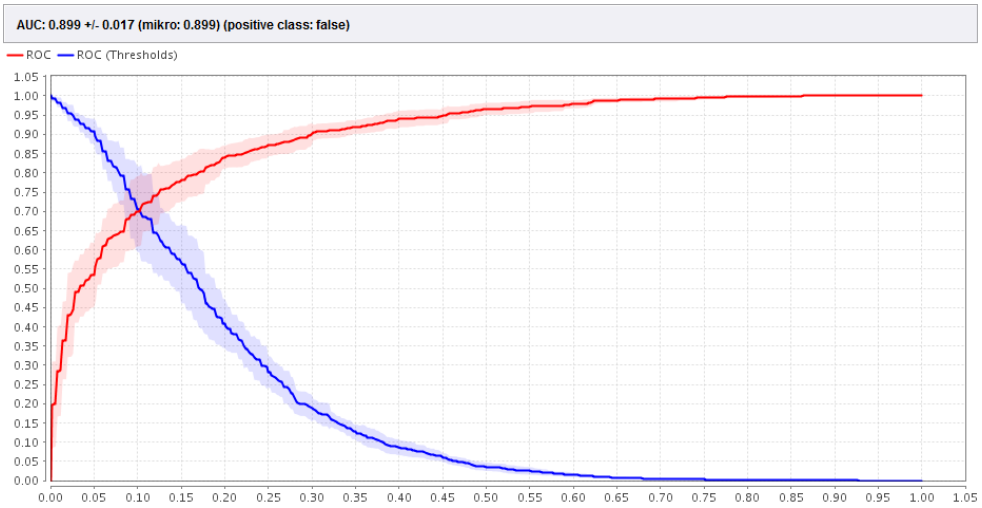

Figure 8.2: AUC Curve for Bayes (Second Model)

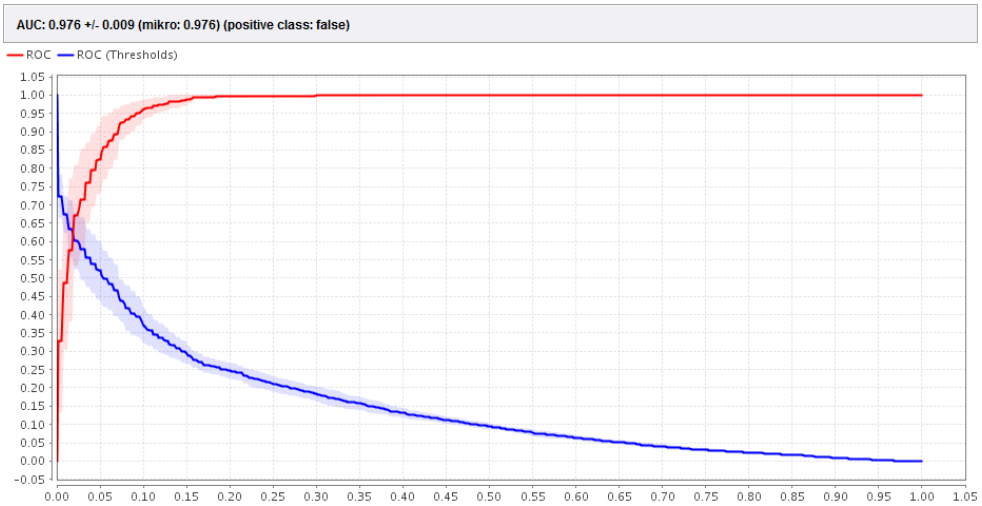

Figure 8.3: AUC Curve for Random Forest (Second Model)

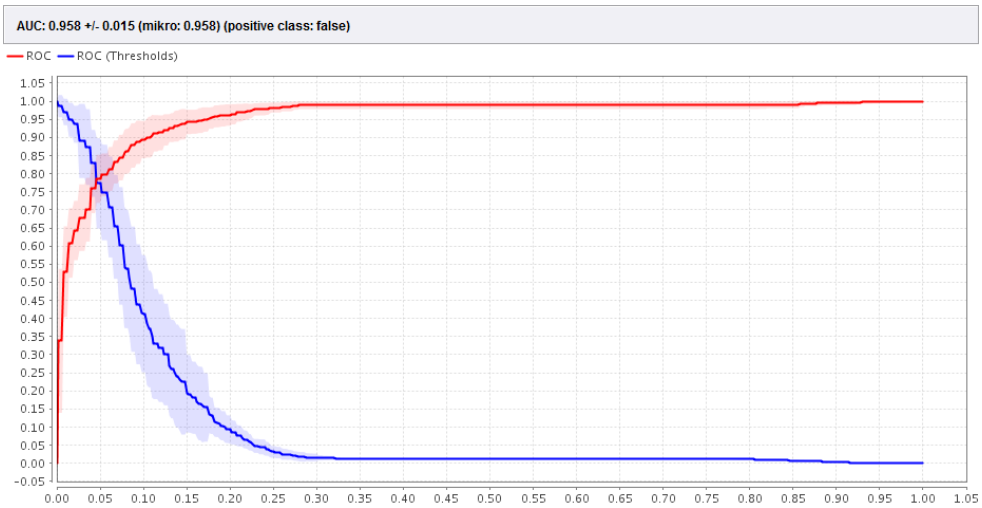

Figure 8.4: AUC Curve for SVM (Second Model)

Appendix B - AUC Curves for Second Model

Chapter 9

Appendix C - Decision Trees

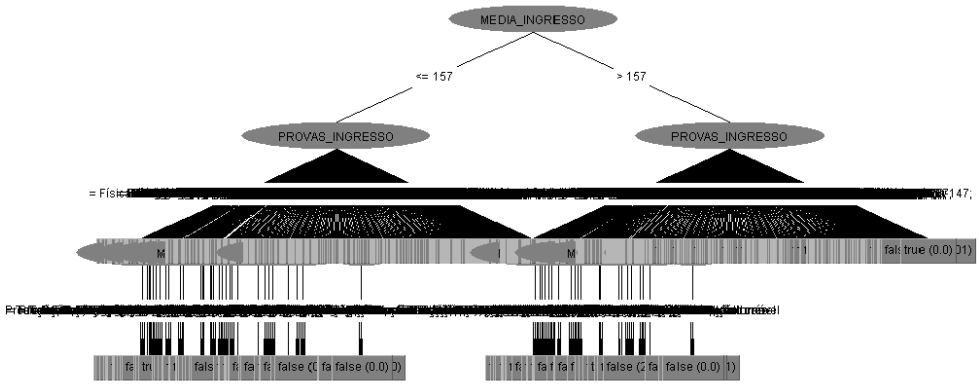

Figure 9.1: Decision Tree of First Model

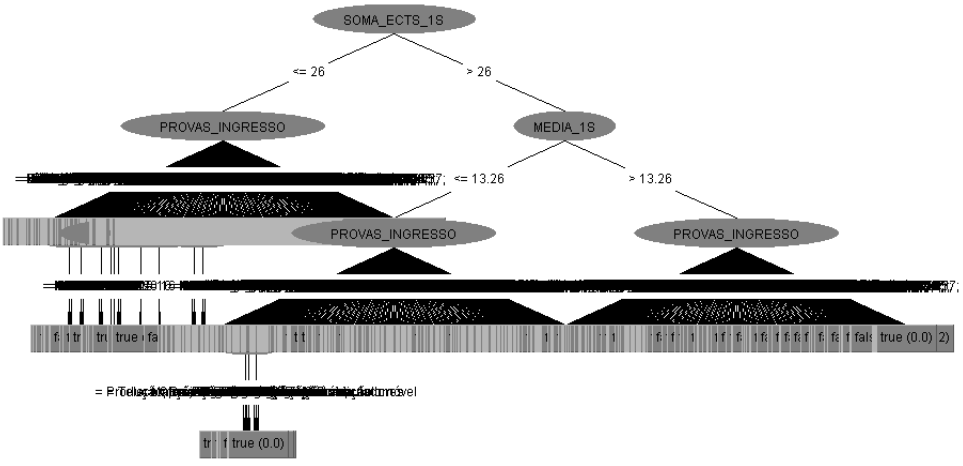

Figure 9.2: Decision Tree of Second Model

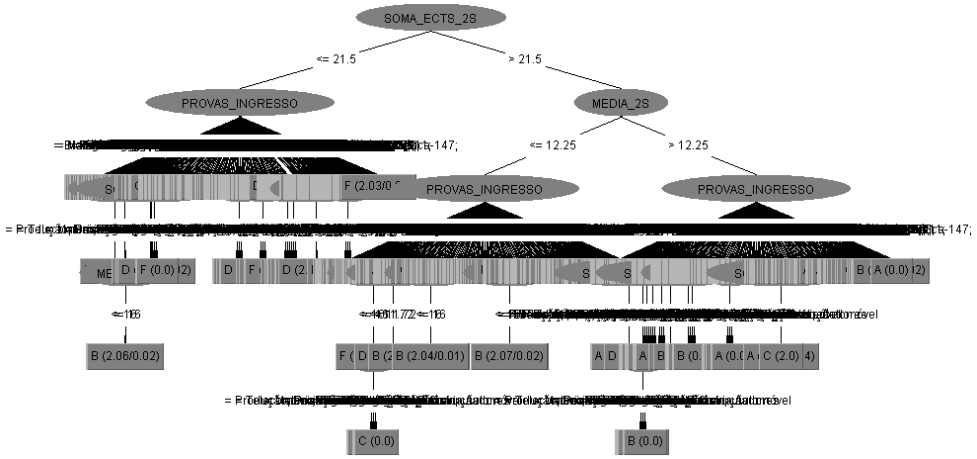

Figure 9.3: Decision Tree of Second Model